AN EXAMINATION OF THE MIMIC METHOD FOR DETECTING DIF AND COMPARISON TO THE IRT LIKELIHOOD RATIO AND WALD TESTS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATIONAL PSYCHOLOGY

MAY 2018

By

Daniel J. Zimbra

Dissertation Committee:

Seongah Im, Chairperson
Min Liu
Yan Wu
George Harrison
Kentaro Hayashi

CONTENTS

TABLES

FIGURES

ABSTRACT

Differential item functioning (DIF) detection research has found the multiple indicators multiple causes (MIMIC) structural equation model (SEM) to be effective in detecting uniform DIF. Recent advances in the MIMIC method have also allowed for the detection of nonuniform DIF. However, few researchers have evaluated its performance, or compared it with the established DIF detection methods. The current study addresses this gap in the existing research by evaluating the MIMIC method in detecting uniform and nonuniform DIF, and comparing its performance to the established item response theory (IRT) based likelihood-ratio (IRTLR) and Wald tests. Monte Carlo simulations of tests and item responses were conducted, manipulating the number of examinees, type of DIF, magnitude of DIF, and proportion of contamination. The simulation results indicate that the MIMIC method outperformed the IRTLR and Wald tests based on Type I error and power rates when testing for a large magnitude of nonuniform DIF and contamination at 20%, regardless of sample size. When the proportion of DIF contamination rose to 40%, the Wald test outperformed IRTLR and MIMIC methods in all other experimental settings. IRTLR was the only method that was able to maintain well-controlled Type I error rates throughout the experimentation and adequate power when the magnitude of DIF was large. While the IRTLR method generally outperformed the others, the MIMIC method was particularly strong at detecting nonuniform DIF, and the Wald test performed well when the proportion of DIF contamination was high. The findings of this study inform future research and practice in the appropriate selection of DIF method.

CHAPTER 1

INTRODUCTION

When using assessments to make meaningful group comparisons (e.g., Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000), researchers and practitioners have become aware of the importance of measurement equivalence across groups (Drasgow, 1984). Differences in test scores across groups that are caused by problems with a measurement instrument rather than true differences in proficiency are concerning to them (Drasgow, 1987; Stark, Chernyshenko, & Drasgow, 2006). Certain methods (i.e., identifiers of differential item functioning or DIF) are necessary to distinguish true differences between test-taker groups. These identify items in assessments that function differently across distinct groups (i.e., gender, ethnicity, age, socio-economic status, language). DIF occurs when a test item displays different measurement properties (e.g., item difficulty; item discrimination) for one group versus another, after taking account of group-mean differences on the test scores (Woods, 2009). Across backgrounds, to ensure validity and fairness of measurement, professionals have developed and refined testing methods and psychometric techniques.

Routinely, the basis for decisions regarding placement, advancement, and licensure, come from test results. Many personal, social, or political implications come from these decisions, so it is crucial the interpretations of a test are valid. Assessment creators attempt to make their measures as accurately as possible to ensure true differences are found between respondents. An item is fair if any person at the same trait level has the same probability of endorsing an item

regardless of their group membership (Woods, 2009). When a test item unfairly favors one group over another, item bias exists, which threatens the test's validity. With item bias, some items function differently, meaning examinees from different groups have unequal probabilities or likelihoods of success on an item, even after they have been matched on the ability of interest (Clauser & Mazor, 1998). For an item to be biased, it is essential that differences exist after matching the ability of interest, because differences in performance alone is not evidence of bias. Performance differences are to be expected when examinees from different groups have different latent ability levels. The result of these differences is called item impact, rather than item bias, meaning the disparity in item performance was the result of legitimate difference in underlying latent factors (Camilli & Shepard, 1994). Overall, ability distributions are reflected by impact (Dorans, Holland, & Wainer, 1993), so when items function differently, there are unexpected differences in performance.

Statement of Purpose

There have been many simulations studies on technical issues of DIF, and as research progresses, many methodological problems appear (Zumbo, 2007). The efficiency and accuracy of DIF methods using item response theory (IRT) and structural equation modeling (SEM) have been reviewed in many of these simulation studies (Cao, Tay, & Liu, 2017; Chun, Stark, Kim, & Chernyshenko, 2016; Finch, 2005; Hou, la Torre, & Nandakumar, 2014; Kristjansson, Aylesworth, Mcdowell, & Zumbo, 2005; Navas-Ara & Gomez-Benito, 2002; Oort, 1998; Raju et al., 2002; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang & Shih, 2010; Wang, Shih, & Yang, 2009; Woods, 2008; Woods, 2009; Woods, Cai, & Wang, 2013). Yet, existing

simulations studies seemed limited in their DIF detection method comparisons, leaving questions about their capabilities. There are important aspects of DIF research that have rarely been addressed or clarified which the present study explores. The purpose for the current research was to offer continued validation and comparison of the detection capabilities of DIF methods utilized in the SEM and IRT. The latest, most effective approaches of DIF detection (i.e., likelihood-ratio test, Wald test, and multiple-indicators multiple-causes SEM) were incorporated to see which one works best in various conditions of item biases.

In previous DIF detection research, although most DIF detection methods had similar rates of correct detection (power) and incorrect detection (Type I error), performance changed when sample size, test length, number of groups, item discrimination, presence and type of DIF, DIF magnitude, and response type (dichotomous/polytomous) were manipulated in various ways (Cao et al., 2017; Chun et al., 2016; Finch, 2005; Hou et al., 2014; Kristjansson et al., 2005; Navas-Ara & Gomez-Benito, 2002; Oort, 1998; Raju et al., 2002; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang & Shih, 2010; Wang et al., 2009; Woods, 2008; Woods, 2009; Woods et al., 2013). Generally, the statistical power to detect DIF increased for each method when there were longer tests, larger sample sizes, or increases in: item discrimination, the number of uncontaminated anchors, the proportion of DIF in the data (Stark et al., 2006). Ideally, the performance of DIF detection procedures should be unaffected by any variation in study conditions (Kristjansson et al., 2005).

Upon reviewing relevant literature, it became apparent that the accuracy of the multiple-indicator multiple-cause (MIMIC) DIF approach (Woods, 2009) has not been completely verified, especially when calculating nonuniform DIF (Woods & Grimm, 2011; Chun et al.,

2016; Lee, Bulut, & Suh, 2017). Originally, MIMIC method was only capable of detecting uniform DIF, but Woods and Grimm (2011) developed a way to calculate nonuniform DIF using MIMIC-interaction models. It is necessary for MIMIC method to accurately calculate nonuniform DIF, a crucial part of DIF detection and research, to compete with methods that can detect both types of DIF. Also, Woods et al. (2013) stated that when MIMIC could accurately test for nonuniform DIF, a comparison of DIF detection performance with the Wald test would be important. No research was found comparing MIMIC with the Wald test.

For that reason, the current study compared the SEM MIMIC approach (Chun et al., 2016) to correctly identify cases of uniform and nonuniform DIF with the IRT based likelihood-ratio tests (IRTLR) and Wald test approaches (Langer, 2008; Woods et al., 2013; Tay, Meade, & Cao, 2015; Cao et al., 2017). Based on recent research (Woods et al., 2013; Tay et al., 2015; Cao et al., 2017; Chun et al., 2016), these methods are the newest and most capable of the DIF detection methods, but need continued comparison in simulation research to verify which works best depending upon testing conditions. Comparing recently developed methods like the Wald test in IRT and MIMIC method in SEM to a more established DIF detection method like IRTLR will help researchers and practitioners decide the optimal method for their purposes.

To accomplish this research, a Monte Carlo study was designed to simulate dichotomous data using IRT item parameters based on the SAT verbal items originally used by Donoghue and Allen (1993) under a variety of manipulated conditions for each method. Parameters that were manipulated in simulation processes included number of examinees in reference and focal groups, type of DIF, DIF magnitude, and proportion of DIF items. These experimental conditions were chosen because they were the most commonly manipulated conditions in

previous research as well as some of the most influential on DIF detection performance (e.g., Cao et al., 2017; Chun et al., 2016; Hou et al., 2014; Kristjansson et al., 2005; Raju et al., 2002; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang & Shih, 2010; Wang et al., 2009; Woods, 2008; Woods, 2009). The simulated conditions in the current research were similar to the aforementioned research for the sake of comparability.

CHAPTER 2

LITERATURE REVIEW

Differential Item Functioning

Test items must operate the same way for participants who are at the same trait level to validly determine whether participants actually differ. DIF occurs when there is an item that examinees with equal trait level, but from different subgroups, do not have an equal probability of endorsing the item positively or answering the item correctly (Hambleton & Rogers, 1989; Hambleton, Swaminathan, & Rogers, 1991; Holland & Wainer, 1993; Lord, 1980; Smith & Reise, 1998; Stark et al., 2006; Steinberg & Thissen, 2006; Swaminathan and Rogers, 1990; Woods, 2008; Woods, 2009; Woods & Grimm, 2011; Woods, Oltmanns, & Turkheimer, 2009). Many DIF detection procedures exist and are described in the literature. Based on theoretical strengths, only a few methods have emerged as preferred after numerous empirical and simulation comparisons. In each of these approaches, there is a comparison of performance on a studied item after matching examinees on the ability of interest. Camilli and Shepard (1994) and Holland and Wainer (1993) gave thorough reviews of some of the methods that exist for identifying DIF, and many since have also studied these methods to decipher their DIF detection capabilities (e.g., Hou et al., 2014; Jiang & Stout, 1998; Kristjansson et al., 2005; Lee et al., 2017; Navas-Ara & Gómez-Benito, 2002; Oort, 1998; Raju et al., 2002; Roussos & Stout, 1996; Stark et al., 2006; Swaminathan and Rogers, 1990; Wang & Shih, 2010; Wang et al., 2009; Woods, 2008; Woods, 2009; Woods et al., 2013). There have been many advances in

methodology, and research continues to compare and validate these methods to decipher the most effective ones.

There are two main categories of DIF detection approaches: 1) the observed summed scores approach and 2) the latent variable approach. Since the 1970s, the observed-score method has been frequently used for investigations detecting DIF. However, in the past decade, the latent variable approach has been of greater interest to researchers. These researchers have focused on studying and improving latent variable methods such as: IRTLR tests (Thissen & Steinberg, 1988), Lord's (1977, 1980) Wald (1943) χ^2 test, the improved Wald test (Langer, 2008; Woods et al., 2013; Tay et al., 2015; Cao et al., 2017), MIMIC model (Jöreskog & Goldberger, 1975; Muthén 1985, 1989; Woods, 2009), MIMIC-interaction model (Woods & Grimm, 2011) with sequential-free baseline (Chun et al., 2016).

The latent variable methods fall under two major approaches, SEM and IRT, which are both capable of detecting item bias and comparing equivalency of measures. They are driven from theoretically different approaches that have separate procedures to examine relationships amongst items and scales, as well as their own terminology. Because of the division between approaches, researchers have investigated and compared both methods to address research questions.

Researchers are essentially interested in two types of DIF that have been identified, uniform and nonuniform DIF (Mellenbergh, 1983). From an IRT framework, which type of DIF occurs depends on the group difference in item parameters, with *a* parameter referring to item discrimination (which is analogous to factor loading in SEM framework), and the *b* parameter referring to item difficulty (which is analogous to threshold in SEM framework).

When *b* parameters differ across groups, uniform DIF is present (Figure 1). The concept of uniform DIF is important in both IRT and SEM, and happens when the probability of

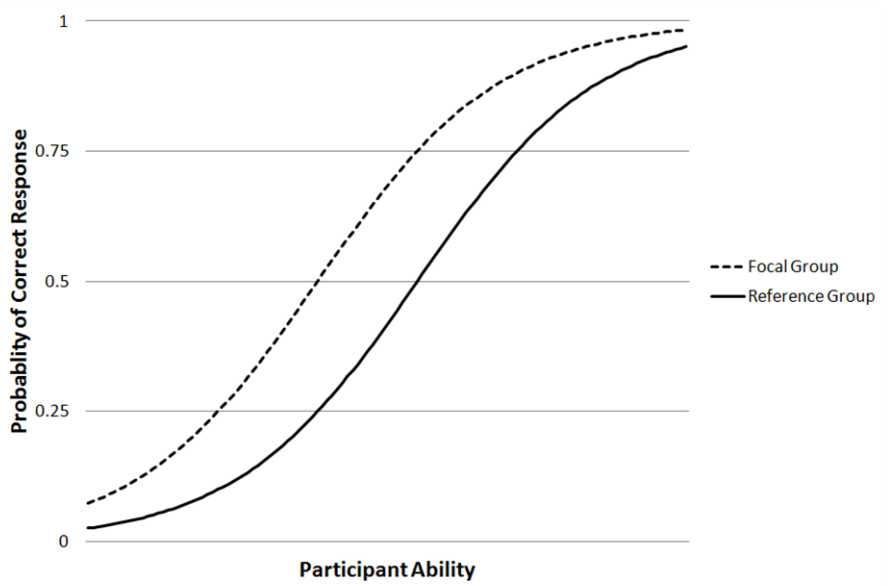


***Figure 1*. Uniform DIF Example**

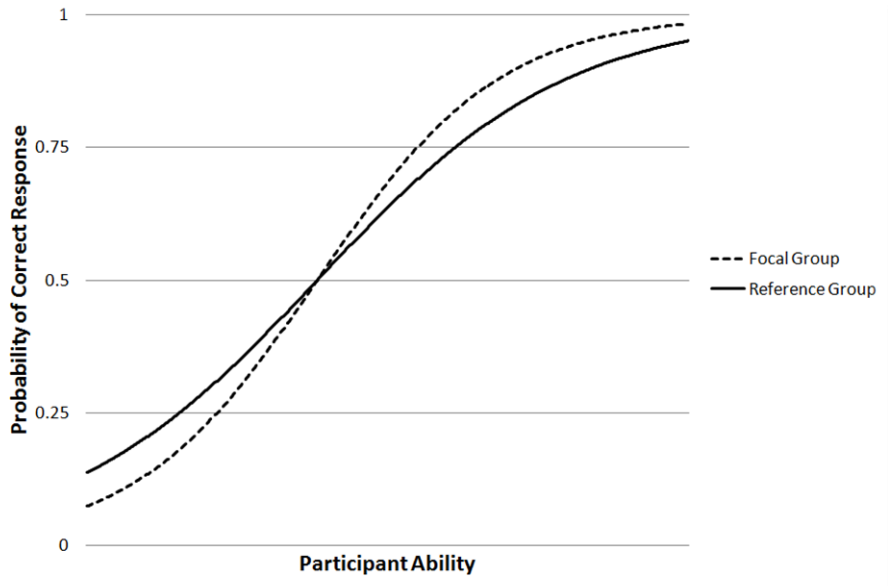


***Figure 2*. Nonuniform DIF Example**

correctly responding to an item is uniformly higher for either the reference or focal groups across ability levels. In the case of uniform DIF, the item characteristic curves (ICCs) of two groups are parallel and do not cross. If, for example, the latent trait represents mathematics ability, then uniform DIF means members of the focal group have a lower probability of answering a mathematics assessment item correctly than the reference group at the same latent trait mathematics ability level. This represents the item being more difficult for one group than for the other along the entire latent construct continuum.

When *a* parameters differ across groups, nonuniform DIF exists (Figure 2). The ICCs of two groups with nonuniform DIF cross each other, meaning group membership and the latent ability level are interacting. With nonuniform DIF, items discriminate differently for the groups, which is why the ICCs are different. Therefore, items with nonuniform DIF are less discriminating for the focal group, and items with uniform DIF are more difficult for the focal group to answer correctly (Woods, 2008).

In the case when uniform and or nonuniform DIF exists, a test item(s) favors one group over another, and therefore, the test is not fair. It is realistic and common to find both uniform and nonuniform DIF during detection processes (French, Hand, Nam, Yen, & Vazquez, 2014). The nature of the DIF (whether it is uniform or nonuniform) will have an effect on the power to detect DIF (Swaminathan & Rogers, 1990), as well as Type I error rates (Kristjansson et al., 2005).

Methods to Detect DIF using IRT and SEM

Several researchers have discussed the connection between IRT and SEM models (Finch, 2005; Fleishman, Spector, & Altman, 2002; Glöckner-Rist & Hoitjink, 2003; MacIntosh & Hashim, 2003; Stark et al., 2006; Takane & de Leeuw, 1987). Both models are capable of estimating latent examinee capability with responses on a test item. They also both provide parameter estimates which describe the items and examinees, including the underlying ability of each examinee, and the item difficulty and discrimination. According to MacIntosh and Hashim (2003), and Muthén, Kao, and Burstein (1991), MIMIC model parameter estimates can be converted to common IRT parameter estimates.

In both IRT and SEM approaches, the analysis begins with a general test for DIF in the discrimination parameter, a_i, or the threshold parameter, b_i. The null (H_0) and alternative (H_a) hypotheses are the following:

H_0: $a_{iF} = a_{iR}$ and $b_{iF} = b_{iR}$

H_a: not all parameters for item i are group invariant

where F is for *focal* group and R is for *reference* group.

The item response function (IRF) of the three-parameter logistic (3PL) model is:

$$P(y_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (1)$$

where θ is the latent trait, c_i is the pseudo-guessing parameter for item i, a_i is the discrimination parameter for item i, and b_i is the difficulty parameter for item i.

In an assessment, an IRF can be estimated for each item, providing the relationship between the probability of producing a correct response and θ. The two-parameter logistic (2PL) IRF does not involve the pseudo-guessing parameter, and is expressed as

$$P(y_i = 1|\theta) = \frac{1}{1+ e^{-1.7a_i(\theta - b_i)}} \quad (2)$$

The IRT parameters are also related to parameters of MIMIC models. Muthén et al. (1991) provided equations to convert the MIMIC parameters to *a* and *b* parameters of two-parameter IRT models as follows:

$$a_i = \lambda_i (1- \lambda_i^2 \psi)^{-1/2} \sigma_\theta \quad (3)$$

$$b_{ik} = [(\tau_i - \beta_i z_k)\lambda_i^{-1} - \mu_\theta]\sigma_\theta^{-1} \quad (4)$$

where z_k represents the indicator for group *k*, in which focal group membership is indicated by *k* equaling 1, and reference group membership by *k* equaling 0; β_i is the estimate of the relationship between the group and item response (where a significant value of β_i indicates the presence of DIF); μ_θ is the mean of the latent trait; ψ is the error variance of the latent trait variable; and σ_θ is the standard deviation of the latent trait. When the latent variable is standardized, the equations for conversion is simplified as μ_θ=0 and σ_θ=1. Muthén et al. (1991) and MacIntosh and Hashim (2003) provide a more complete explanation of these relationships. Due to the high correspondence between these two approaches, it is reasonable to apply one to answer questions raised by another, as well as compare performance and effectiveness when using them to examine DIF.

DIF Approaches in SEM and IRT

SEM MIMIC Method

The MIMIC model was first described by Jöreskog and Goldberger (1975) as a special application of SEM. Muthén (1988, 1989) then applied this method in research, and more fully articulated the method to investigate latent variable modeling in heterogeneous groups and

examine potential group effects on both latent and observed variables. The MIMIC model is critical to validation research because it allows the investigation of multi-group differences on a latent construct (Hancock, 2001). It can be used to examine potential DIF in the observed indicators of the latent variables (Muthén, 1989), and is efficient in handling heterogeneity in populations. It can be used to (1) assess a test's construct validity by fitting a theoretical model to a set of data via confirmatory factor analysis (CFA), (2) determine if latent factor means differ between populations, and (3) examine the measures of the latent factors for potential DIF (Muthén, 1988). The standard CFA model is extended in the MIMIC model by including exogenous variables that affect the latent factors, creating one data set with all combinations of populations of interest. It involves unobserved latent factors caused by several z-variables (covariates), and is indicated by several y-variables. The model equations are:

$$y = \lambda\theta + \varepsilon \tag{5}$$

$$\theta = \gamma' z + \zeta \tag{6}$$

where $y' = (y_1, y_2, \ldots y_k)$ are the indicators of the latent variable θ, and $z' = (z_1, z_2, \ldots z_k)$ are the causes of θ. From these equations, we have:

$$y = \lambda\gamma' z + \lambda\zeta + \varepsilon \tag{7}$$

$$= \Pi z + w \tag{8}$$

Thus, $\Pi = \lambda\gamma'$, $w = \lambda\zeta + \varepsilon$ and $Cov(w) = \lambda\lambda'\psi + \Theta_\varepsilon$, where $\psi = \mathrm{Var}(\zeta)$, and Θ_ε is the diagonal covariance matrix of ε (Jöreskog & Sörbom, 2002).

A latent response variable formulation from Muthén and Asparouhov (2002) is used to test for DIF using the MIMIC model for dichotomous items.

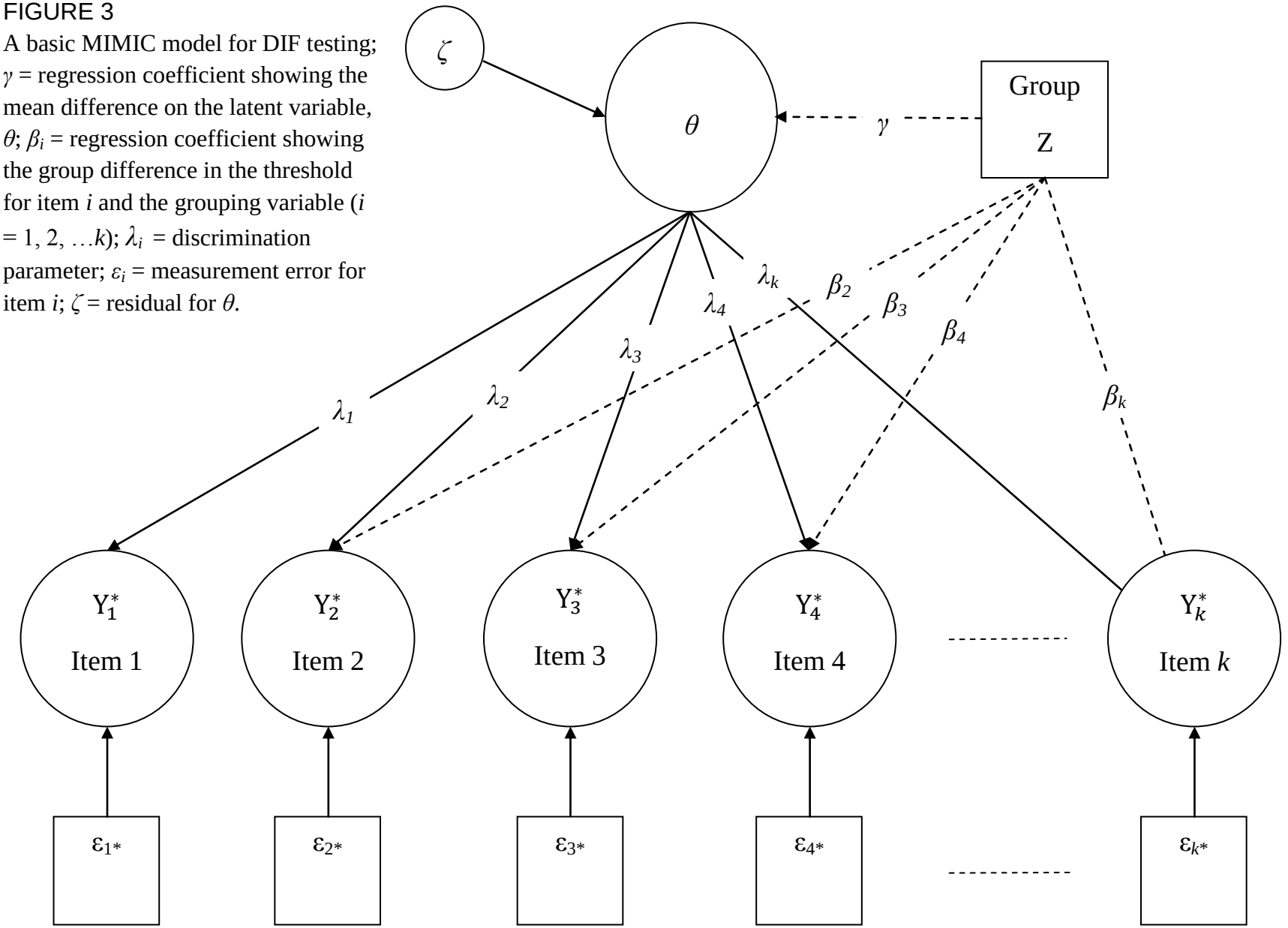
The MIMIC model in the DIF context is:

$$y_i^* = \lambda_i\theta + \beta_i z_k + \varepsilon_i \quad (9)$$

where y_i^*, λ_i, θ, and ε_i are as defined above; z_k = a dummy variable indicating group k membership; and β_i = slope relating the group variable with the response.

FIGURE 3

A basic MIMIC model for DIF testing; γ = regression coefficient showing the mean difference on the latent variable, θ; β_i = regression coefficient showing the group difference in the threshold for item *i* and the grouping variable (*i* = 1, 2, …*k*); λ_i = discrimination parameter; ε_i = measurement error for item *i*; ζ = residual for θ.



The MIMIC-interaction model from Woods and Grimm (2011) is capable of testing uniform and nonuniform DIF simultaneously in a dichotomous item and is based on the continuous latent response variable:

$$y_i^* = \lambda_i\theta + \beta_i z_k + \omega_i\theta z_k + \varepsilon_i \quad (10)$$

where λ_i is the factor loading of item i on the latent variable θ, β_i indicates the uniform DIF effect or direct effect (when $\beta_i \neq 0$) showing the group difference in the threshold parameter after controlling for any mean ability difference on θ between groups, z_k is the categorical covariate (where $k = 0$ for the reference group and $k = 1$ for the focal group), ω_i is the interaction term between the latent trait and the categorical covariate (i.e., group variable) that represents the nonuniform DIF effect (when $\omega_i \neq 0$), and ε_i is the error term that is normally distributed and independent of θ and z (Lee et al., 2017). The MIMIC-interaction model is very similar to the DIF detection in the logistic regression approach by Swaminathan and Rogers (1990). After controlling for group differences in the latent trait, both approaches test the difference in the probability of answering a dichotomous item correctly due to group membership and the interaction between the group membership and the latent trait (Lee et al., 2017).

With the goal of determining whether items measuring a latent variable are equally discriminate and difficult across comparison groups, MIMIC DIF analysis involves comparing the fit of a series of full and reduced models. MIMIC is unique because rather than fixing and freeing parameters reflecting item loadings (discrimination) and thresholds (difficulty) across groups, MIMIC tests for DIF by adding or deleting direct paths to items emanating from the background variables associated with group membership, and impact is accounted for by paths from grouping variables to the common factor. Essentially, MIMIC tells us how grouping variables affect item properties and factor means.

For convenience, just one set of model parameters using the total sample of participants is estimated in the MIMIC methods. This allows MIMIC to effectively test for DIF using smaller sample sizes because the full sample (rather than two separate groups) is used for estimation (Muthén, 1988, 1989). Power to detect true heterogeneity in the original population can also increase by keeping the sample as a whole. Also, when more groups are compared, sample size does not need to increase. MIMIC models are an attractive alternative because they can investigate why DIF occurs by allowing the inclusion of more than one background variable and its interactions.

The MIMIC model has been applied to detect DIF in various areas of study, and during the review of the literature from 1990 on, 19 simulation studies were found investigating its performance. As an example of one of these studies, Finch (2005) compared the uniform DIF detection of three methods (Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and IRTLR) with MIMIC and found: 1) MIMIC had higher Type I error rate than any other method in a short test of 20 items using data generated from the 3PL IRT model; 2) when using data from the 2PL IRT model, MIMIC had equivalent Type I error as the other three methods when the test was longer (50 items).

Hong, Wu, Maller, & Pei (2008) performed a Monte Carlo simulation study to detect DIF using MIMIC, checking for both uniform and nonuniform DIF. They found that: 1) when the studied item had low discrimination, MIMIC could not detect DIF due to impaired power; 2) Type I error was fair; 3) low Type I error rate was reported for MIMIC even with smaller focal group (vs. reference group) sample size, giving confirmation that MIMIC performs well by incorporating background variables into the model. MIMIC method has this advantage over

other DIF detection methods; and 4) under certain conditions, MIMIC has strong power to detect uniform DIF.

Using simulated data, Wang et al. (2009) compared the standard MIMIC method to the MIMIC method with scale purification procedure to detect DIF and assess performance. A scale purification procedure consists of four steps: (a) initially estimating DIF in all items, (b) removing items detected as having DIF from the set of anchor items that do not have DIF, (c) re-estimating DIF in all items, and (d) repeating steps b and c until the same set of items are detected as having DIF in two consecutive iterations (Wang et al., 2009). The standard MIMIC method was outperformed by the MIMIC method with scale purification. The authors suggested that the MIMIC method with scale purification is preferable because DIF patterns in real tests are unlikely to be perfectly balanced (with equal amounts of uniform and nonuniform DIF) and the percentages of DIF items may not be small. In a series of simulations using polytomous items, Wang and Shih (2010) used three MIMIC methods (standard, with scale purification, with pure anchor) to assess DIF. In the MIMIC with pure anchor method, anchor items are preselected via the scale purification process described above and are constrained to be equal across groups, and the parameters of other items are allowed to differ across groups. It was shown that MIMIC method with a pure anchor set yielded very high accuracy in comparison to the other approaches, and maintained Type I error rate and high power even when test contained as many as 40% DIF items.

In a simulation study, Woods (2009) examined the Type I error and power rates of the MIMIC and IRTLR methods when detecting DIF with small focal groups. Results indicated that the MIMIC approach performed better than IRTLR in testing for uniform DIF with small focal groups. In a similar simulation study comparing MIMIC and IRTLR methods to detect DIF,

Woods and Grimm (2011) added a latent variable interaction to the MIMIC model to test for nonuniform DIF using MIMIC for the first time. The main finding was that when the latent moderated structural equations approach was used to estimate the interaction to test for nonuniform DIF, the Type I error in the MIMIC model with the interaction was inflated. Lee et al. (2017) extended the research of Woods and Grimm (2011) by examining the performance of a multi-dimensional MIMIC-interaction model under various simulation conditions with respect to Type I error and power rates. The study concluded that power rates were higher in the uniform DIF conditions than in the nonuniform DIF conditions. Also, power to detect DIF increased with larger sample sizes and more anchor items. Overall, the multidimensional MIMIC-interaction model was sufficient at detecting uniform DIF, but nonuniform DIF detection capabilities were still questionable.

Chun et al. (2016) conducted a simulation study to investigate the efficacy of MIMIC methods for multi-group uniform and nonuniform DIF. DIF was simulated to originate from two background variables (i.e., gender and ethnicity) and three implementations of MIMIC DIF methods were compared: constrained baseline, free baseline, sequential-free baseline. Most MIMIC DIF research has been conducted using the constrained baseline method in which items are tested for uniform DIF, associated with group differences in item thresholds, by adding paths from each grouping variable to individual items in a sequence of reduced versus full model comparisons. If a full model fits significantly better than the reduced baseline, then the item under investigation is flagged as a DIF item (Kim et al., 2012). Although the constrained baseline approach to testing for DIF is convenient because it allows every item to be evaluated, it often leads to high Type I error rates because the baseline model is incorrectly specified when DIF is present (Stark et al., 2006). Free-baseline approaches to DIF detection begin by forming a

baseline model that has only the necessary constraints for identification. This is accomplished by constraining the loadings and thresholds for one item to be equal across comparison groups. Reduced models are then formed by constraining the loading and threshold parameters simultaneously for one additional item at a time and examining the change in goodness of fit for each reduced model relative to the baseline. For the sequential-free baseline test depicted in Figures 4 and 5, first, conduct constrained baseline tests to identify items that appear to be free of DIF. Then, choose the most discriminating non-DIF item as the anchor for subsequent free baseline tests of the other items in the scale. If the fit worsens significantly, then the item under investigation is flagged as DIF. The sequential-free baseline approach outperformed the other implementations, providing excellent Type I error and power.

Researchers have been interested in testing DIF in several groups simultaneously to identify bias within assessments. MIMIC method is an effective alternative to traditional IRT methods because it can easily accommodate background variables and their interactions without needing large samples (Kim, Yoon, & Lee, 2012; Woods, 2009). Also, SEM software advances that allow for interactions between latent and observed variables make it possible to detect uniform *and* nonuniform DIF (Woods & Grimm, 2011; Lee et al., 2017). To avoid statistical corrections for contamination due to DIF in constrained baseline applications, free baseline

FIGURE 4

A MIMIC model for testing nonuniform DIF with interaction between group and θ with Item 1 as designated anchor; γ = mean difference on the latent variable, θ; items i = 1, 2, …, k; λ_i = loading; ω_i = nonuniform DIF effect; τ_i = threshold; β_i = group difference in the threshold.

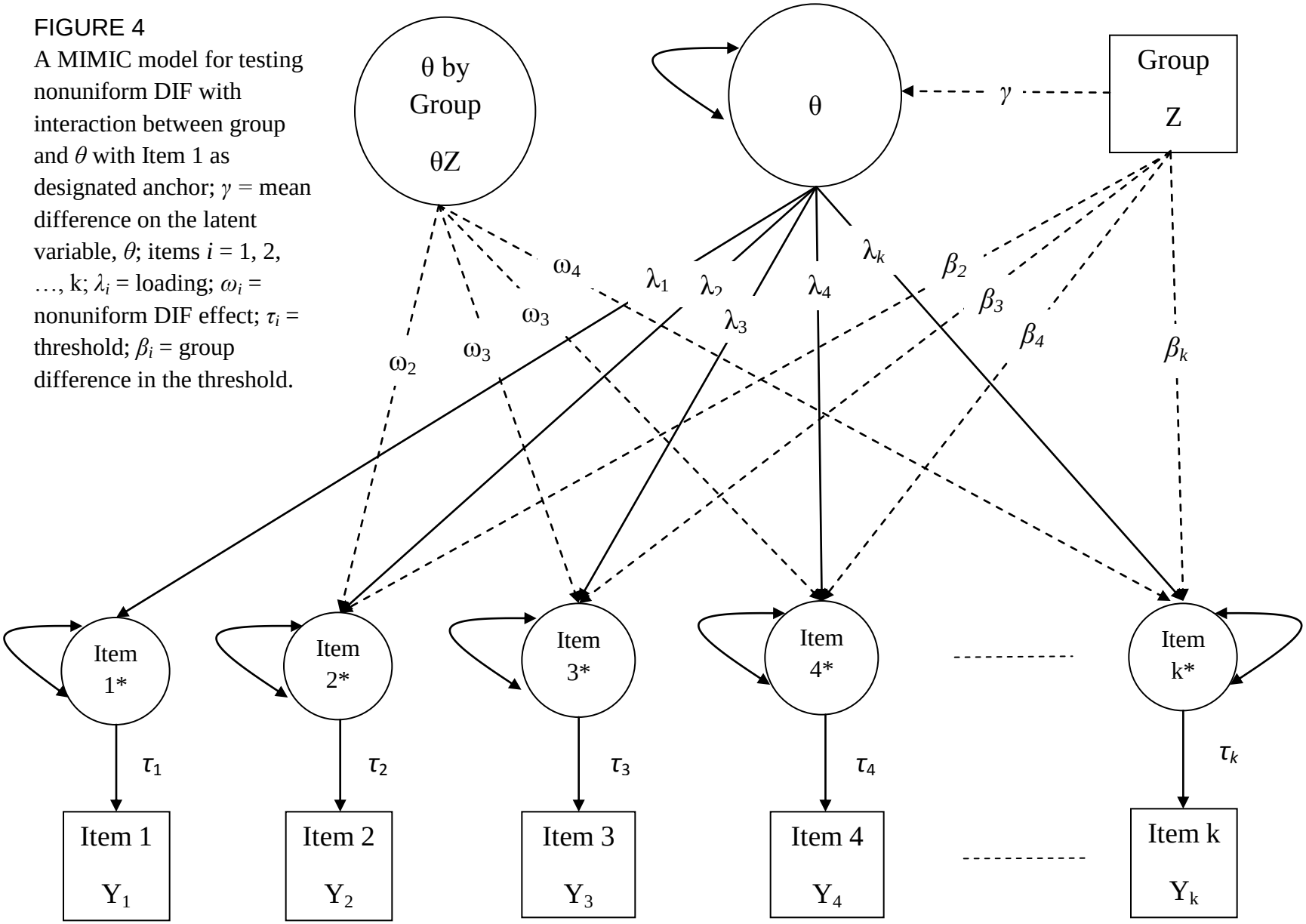
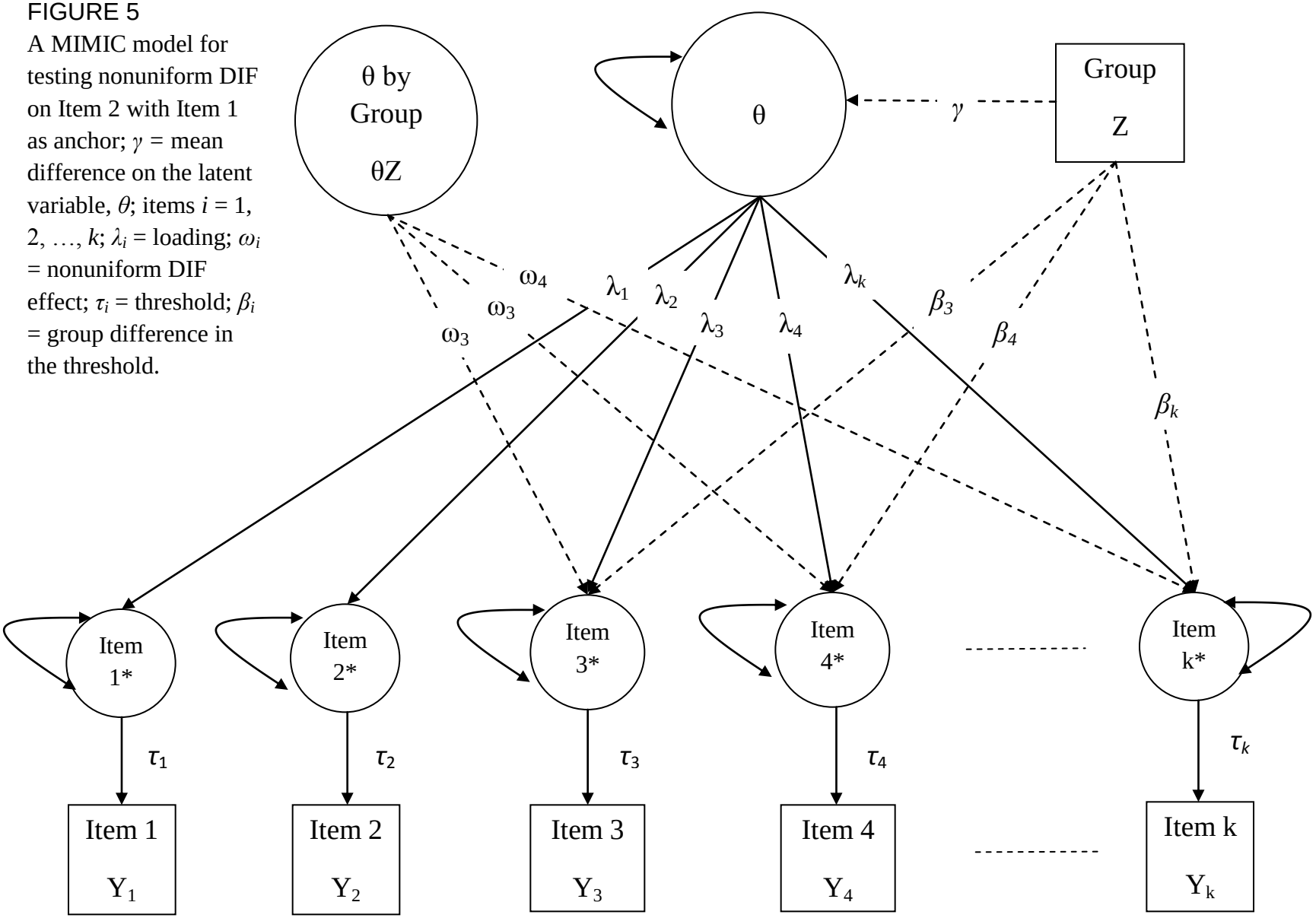


FIGURE 5

A MIMIC model for testing nonuniform DIF on Item 2 with Item 1 as anchor; γ = mean difference on the latent variable, θ; items i = 1, 2, …, k; λ_i = loading; ω_i = nonuniform DIF effect; τ_i = threshold; β_i = group difference in the threshold.



MIMIC methods for DIF analysis can be used and provide Type I error control and high power (Chun et al., 2016).

IRT Likelihood Ratio Test

The IRTLR test is another established approach to DIF detection that has been compared with other methods in previous simulation research. Originally described by Thissen, Steinberg, and Gerrard (1986) and later expanded by Thissen, Steinberg, and Wainer (1993) for dichotomous and polytomous data, this method allows for model fit comparison, assuming parameter estimate equality for the item in question across reference and focal groups (compact model), with the model fit when this constraint is relaxed and differences for the item parameters across groups are allowed (free or full model). The IRTLR test detects DIF items by comparing the relative goodness of fit of the two models using the difference in -2 times the log likelihood values of the two models and the test statistic follows a chi-square distribution with the degrees of freedom equal to the difference in the number of parameter estimates in each model. Some items are constrained to be equal for parameter estimates for both groups, referred to as anchor items, in both models. The test statistic takes the following form when comparing two groups on all item parameters simultaneously:

$$LR = (-2\ln L_R) - (-2\ln L_F) \tag{11}$$

where $\ln L_R$ is the log likelihood of the compact model (i.e., more equal condition), and $\ln L_F$ is the log likelihood of the full or free model (i.e., more free conditions).

To test for the presence of DIF in the 2PL model, find the difference between the two models' log likelihood values distributed as a χ^2 statistic with $df = 2$. If the test statistic value is statistically significant, subsequent tests compare the fit of the models to the two groups, with all item parameters except for one held equal. The LR statistic is calculated with the compact

model holding the two parameters equal while the discrimination parameter is allowed to vary in the free model, formulized as:

$$LR = -2\ln L_R - (-2\ln L_{Fa}) \quad (12)$$

where L_{Fa} = log likelihood of the full model, with the discrimination parameter allowed to vary between groups. The same process is undertaken to calculate the LR statistic for the difficulty parameters across the two groups, formulized as:

$$LR = -2\ln L_R - (-2\ln L_{Fb}) \quad (13)$$

where L_{Fb} is the log likelihood of the free model with the *b* parameter allowed to vary between groups.

To test for the presence of DIF, find the difference between the two models' log likelihood values distributed as a χ^2 statistic with $df = 1$. IRTLRDIF (Thissen, 2001) is a software program that has formalized this methodology, but other software programs like the *mirt* package in R (Chalmers, 2012) can perform IRTLR as well.

Simulation research on IRTLR DIF detection has shown that under various realistic conditions, the Type I error for the general IRTLR test is close to nominal level and group-mean difference is recovered when latent variables are actually normally distributed (i.e., assumptions for the test are met) (Ankenmann, Witt, & Dunbar, 1999; Bolt, 2002; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Stark et al., 2006; Sweeney, 1997; Wang & Yeh, 2003; Woods, 2009). Research by Cohen et al. (1996) found that for 3PL models with samples of both 250 and 1,000 examinees and a test of 50 items, Type I error rate for the LR statistic was above nominal 0.05. Wanichtanom (2001) expanded this study to include power, and found that it had a 0.97 detection rate for uniform DIF when averaged across group differences in the *b* parameters (set at 0.2, 0.5, and 0.8).

In IRTLR, if there is an increase in sample size, item discrimination, the number of anchors, or the amount DIF in the data, the statistical power to detect uniform and nonuniform DIF increases (Ankenmann et al., 1999; Wang & Yeh, 2003; Woods, 2009). These results are based on item responses simulated from the 2PL, 3PL, or graded IRFs, with test lengths of 10, 15, 20, 25, 26, 30, 40, or 50 items and group-mean differences of 0, .4, .5, or 1 SD. Sample sizes in these studies were equal for both reference and focal groups, or larger for the reference group. Anchors have been all other items, one item, or 10%, 16%, 20%, or 40% of the total number of items.

IRT Wald Test

The Lord's Wald Test (1977, 1980) compares IRT item parameter vectors between groups using the following formula:

$$\chi^2_i = v^T_i \Sigma_i^{-1} v_i \quad (14)$$

In a 2PL model, when two groups are compared, Σ_i^{-1} represents the estimate of the inverse of sampling variance-covariance matrix of the differences between the item parameter estimates, and $v^T_i = [a_{F_i} - a_{R_i}, b_{F_i} - b_{R_i}]$ (where F represents the focal group and R represents the reference group). To create a χ^2 statistic, the Wald test statistic uses information from the covariance matrix of the differences of item parameter estimates between groups and the actual values of these differences themselves. Although similar to IRTLR (Thissen et al., 1993), the Wald Test performed poorly during DIF detection simulations (Woods et al., 2013). Inflations in Type I error and inaccurate estimation of the covariance matrix were shown after testing (Donoghue & Isham, 1998; Kim, Cohen, & Kim, 1994; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987). For these reasons, publication of methodological work on the Wald test ceased until it was improved (Cai et al., 2011; Langer, 2008; Woods et al., 2013).

Improved Wald Test

As an alternative for the original Wald test (Lord, 1980), a two-stage equating approach was proposed by Langer (2008) to improve the functionality. Type I error was well-controlled in this study, suggesting the proposed Wald test performed better with these improvements. The main changes Langer (2008) made to the original Wald test were the linking/equating procedure and estimation method. To change the linking/equating procedure, the Stocking-Lord (1983) approach was replaced by concurrent calibration (Kolen & Brennan, 2004; Langer, 2008). Rather than ad hoc equating, item parameter estimation was now performed where the latent scale is held constant over groups simultaneously. Regarding accuracy of estimated scores and stability of recovering item parameters, concurrent calibration outperformed the Stocking-Lord (1983) approach, thus making it a logical choice to improve the Wald test (Kim & Cohen, 2002; Petersen, Cook, & Stocking, 1983; Tian, 2011). For the estimation method, Langer (2008) chose supplemented expectation maximization algorithm (Cai, 2008; Meng & Rubin, 1991) to estimate the covariance matrix, further improving performance of the original Wald test. The algorithm is convenient for estimating the information matrix for item parameters, therefore allowing more accurate standard errors for the estimated item parameters (Langer, 2008). Langer's (2008) approach was called Wald-2 by Woods et al. (2013).

More recently, researchers proposed a one-stage equating approach called Wald-1 (Cai et al., 2011; Woods et al., 2013) as an extension to the improved Wald test proposed by Langer (2008). Wald-2 and Wald-1 are the two equating algorithms for the improved Wald test, and share certain statistical characteristics. For example, following the improvements made by Langer (2008), both improve on ad hoc linking by linking the metric across groups simultaneously with parameter estimation. Also, both use supplemented expectation

maximization algorithm (Cai, 2012; Meng & Rubin, 1991) as well as are implemented with IRTPRO (Cai et al., 2011), flexMIRT (Cai, 2013), and R version 3.4.2 (R Core Team, 2017). Expanding upon Langer's (2008) research, Woods et al. (2013) replicated the simulation study and demonstrated utility of Wald-1 for multi-group DIF detection. For example, Woods et al. (2013) found that the improved Wald test performs well when there are unequal sample sizes between groups. This is valuable because it is quite common to have unequal sample sizes in realistic settings. Also, both studies compare IRTLR with the improved Wald test to detect DIF in ordinal responses (Langer, 2008; Woods et al., 2013). In both, IRTLR was outperformed by the improved Wald test.

Wald-1 and Wald-2 have distinct requirements for anchor items due to the employment of different equating algorithms. To assess candidate items and estimate group differences, Wald-1 uses designated anchors to detect DIF. Langer (2008), in contrast, detected DIF using Wald-2 with all items as anchors, and therefore designated anchors were not required. According to the results of Woods et al. (2013), Wald-1 was better at estimating latent means and their variability, outperforming Wald-2. As a limitation, however, Wald-1 cannot specify which item parameters actual differ between groups (Woods et al., 2013).

One of the many advantages of Wald-2 test is that all items are tested for DIF. Also, no anchors need to be specified to use Wald-2. However, Wald-2 does have some disadvantages. For example, because no anchors need to be specified, none are designated. Without designated anchors, if there is DIF that does not cancel out across items in the first testing stage, the focal group mean and standard deviation are estimated from an incorrect model. Type I error may be inflated due to this misspecification. Other inaccuracies may be possible as well (Langer, 2008; Woods et al. 2013).

Similar to most other DIF procedures, the improved Wald-1 test (Cai et al., 2011) requires the designation of anchor items. Prior research and testing using Wald-2 or other methods can dictate anchor item specification. When there are studied items that function differently, empirical anchor selection methods are commonly discussed in the DIF literature, leading to more accurate Type I error rates (Kim & Cohen, 1995; Woods, 2009). Although Wald-1 approach performs better, it requires prior knowledge of anchor items, so it is difficult to implement when anchor items are unknown (Cao et al., 2017).

After the Wald test was improved, simulations of the Wald-2 approach performed by Langer (2008) using a graded response model (Samejima, 1968; Samejima, 1997) with ordinal response items showed Type I error being well-controlled. However, Woods et al. (2013) stated that this was probably because power was quite low. Twenty percent of items on all simulated tests functioned differently between groups, and items were discrepant between groups by a difference of 0.1 or 0.2 in threshold and a multiple of 1.25 or 0.875 in discrimination. Sample sizes were equal for the reference and focal groups, N = 250 or 1,000, and the focal group mean was 0 or 2.6. Woods et al. (2013) used data that was generated with larger differences between groups so that Type I error for Wald-2 would receive a more realistic evaluation. The Wald-2 approach led to unacceptably high Type I error rates in almost all DIF detection conditions (Woods et al., 2013). If the assumption that there is no DIF at the scale level under which the Wald-2 approach estimates the latent trait distribution is not met, then the latent trait estimation of the focal group will likely be biased, leading to inaccurate DIF detection (Tay et al., 2015). Based on simulation results, Woods et al. (2013) recommended DIF detection using the Wald-1 approach (Cai et al., 2011), as it demonstrated superior performance over the Wald-2 approach in terms of Type I error rate and power.

Iterative Wald Test

The iterative Wald test was proposed by Tay et al. (2015) to effectively test for DIF without prior knowledge of anchor items. Analogous to iterative linking approach by Stocking & Lord (1983), this approach puts the reference and focal groups on the same scale by using non-DIF items as anchor items. An iterative procedure is used to further refine these anchors items. Cao et al. (2017) provided four steps to implement the iterative approach which are guided by known advantages of the Wald-1 and Wald-2 tests. For example, Wald-2 has high Type I error rates in the presence of DIF in the other items. Therefore, there is high probability that non-DIF items are identified as having DIF. Anchor items are found by Wald-2 because the approach has good power, so when items do not have DIF, there is more confidence towards the likelihood of these items actually being non-DIF. This parallels the fully-constrained baseline approach used by Stark et al. (2006), where using all items for linking led to high Type I error rates, and subsequently non-DIF items are suggested for use as anchor items. Also, when anchor items are known, the Wald-1 approach is known to have considerable power and good Type I error rates. During the test, any item that does not display DIF is assumed to be an anchor because Wald-1 has well-controlled Type I error rates. Based on the Wald χ^2 test, when all non-anchor items display significant DIF, the procedure is complete.

Only two articles (Tay et al., 2015; Cao et al., 2017) illustrated the use of the iterative Wald test approach to detect DIF. Analyzing both dichotomous and polytomous data sets, the iterative Wald test was successful in detecting DIF in both studies. However, being only based on two simulated samples, these studies were unable to provide information about the performance of the iterative Wald approach to detecting DIF under a variety of simulation

conditions. According to Cao et al. (2017), the iterative Wald test needs to be rigorously examined using Monte Carlo simulations.

Strengths and Weaknesses of the Three DIF Methods

According to research on the MIMIC model, it has several advantages. First, the framework of the MIMIC model is flexible to accommodate any group variable with two or more levels. With MIMIC, it is possible to check for DIF in more than two groups, with multiple categorical or continuous background variables (Glöckner-Rist & Hoitjink, 2003), and offers a more complete examination of how they relate to the latent trait (Muthén, 1988). Other methods cannot accomplish this, making it more difficult to examine differences when more than two groups are present. Only two groups can be compared at a time (as in some advanced IRT methods as well), but MIMIC model can specify all groups using covariates and examine simultaneously.

Second, because the factor loading matrix is assumed to be identical across populations in the MIMIC model, it requires less parameter estimation than other more complex IRT DIF detection methods. However, this assumption may also be a concern when using MIMIC method (Hong, 2010).

Third, MIMIC model receives extra information from important background variables, enabling researchers to investigate construct validity and invariance hypotheses across sub-populations (Muthén, 1988). With regular factor analysis, the difference in background groups may not be captured due to the covariance matrix only containing response variables (Muthén, 1989).

Finally, MIMIC has performed well in dichotomous (Finch, 2005; Stark et al., 2006) and polytomous (Chun et al., 2016; Hong et al., 2008; Wang & Shih, 2010) items, and can

investigate potential DIF for each item through specification and estimation. Each covariate may have differential effects, and can be checked through examining the statistical significance of the direct path from the covariate to the item (Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000). A significant direct path from group to item suggests potential DIF, and may prevent accurate comparison of the latent factor means. The key limitation for MIMIC relative to other DIF detection methods, however, is its ability to examine just two potential sources of invariance (i.e., indicator intercepts, factor means). Regardless, with many advantages, MIMIC is proven worthy of further investigation regarding its DIF detection capabilities.

Previous research has shown that DIF detection using IRTLR approach has more power than other methods (Thissen et al., 1993; Teresi, Kleinman, & Ocepek-Welikson, 2000; Wainer, 1995; Wanichtanom, 2001). Various factor manipulations increase this power, including number of anchors, sample size, and item discrimination (Wang & Yeh, 2003). With good power, IRTLR tests have minimal chance of accepting the null hypothesis of no DIF, further supporting this method as one of the most powerful in detecting DIF. Also, in simulation research, Type I error for IRTLR test was close to nominal level and group-mean difference was recovered when assumptions for the test were met (Cohen et al., 1996; Stark et al., 2006; Woods, 2009).

However, one of the main disadvantages of IRTLR is that with each additional set of hypotheses under examination, there is an increase in the number of models needing to be fit. It is necessary that the compact as well as augmented models both be fitted to the model, so fitting the model happens twice per hypothesis. Plenty of model fittings and computational time is required, especially with more than two groups. If researchers do not want to use IRTLR due to these disadvantages, yet want to use a similar method that easily extends to multiple groups,

Lord's Wald test (1977, 1980) is asymptotically equivalent (Thissen et al., 1993; Teresi et al., 2000; Wainer, 1995).

The Wald test has many advantages. For example, simultaneous employment of DIF detection across several groups makes the Wald test more efficient. Furthermore, the Wald test is better at detecting DIF not only in uniform DIF, but also nonuniform DIF (Woods et al., 2013). In comparison, the MIMIC-interaction model can detect nonuniform DIF, but this has yet to be thoroughly investigated. Also, multi-group comparisons are easily accomplished with software programs such as IRTPRO (Cai et al., 2011) and flexMIRT (Cai, 2013), and the *mirt* package in R (Chalmers, 2012).

However, the Wald test does suffer from limitations. For example, the test may require a larger sample size in comparison to other tests. Also, in theory, the test can accommodate covariates (e.g., other factors such as demographics) that can be statistically controlled besides θ, but at the moment, specification of covariates is not allowed by IRTPRO (Cai et al., 2011) and flexMIRT (Cai, 2013).

Evaluation Criteria and Anchor Item Selection

Quality of performance was based on measures of power and Type I error as in previous DIF detection simulation comparison research (Cao et al., 2017; Chun et al., 2016; Finch, 2005; Hou et al., 2014; Kim et al., 2012; Kristjansson et al., 2005; Lee et al., 2017; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang et al., 2009; Woods, 2008; Woods, 2009; Woods et al., 2009; Woods & Grimm, 2011; Woods et al., 2013). In the current research design, measuring these variables allowed the evaluation of performance capabilities of the methods and subsequently to compare them.

Type I Error

Type I error rates have been shown to vary according to method used and simulation conditions. Type I error rate is defined as the percentage of time the item was detected as displaying DIF out of the amount of replications under each non-DIF condition (Hou et al., 2014). It is also the probability of rejecting a true null hypothesis. Inflated Type I error means some DIF-free items will appear to have DIF and not be selected for the DIF-free subset (Woods et al., 2009). For example, the assumption that all other items are DIF-free is increasingly incorrect for scales with more DIF. Previous simulation studies indicate that the error produced by violation of this assumption is inflated Type I error (Finch, 2005; Stark et al., 2006; Wang, 2004; Wang & Yeh, 2003).

Several researchers have reported that DIF detection procedures exhibit high Type I error when groups differ in average ability and when the studied item's discrimination is high (Chang, Mazzeo, & Roussos, 1996; Tian, 1999; Zwick, Thayer, & Mazzeo, 1997). For example, Chang et al. (1996), Spray and Miller (1994), Tian (1999), and Zwick et al. (1997) reported that polytomous DIF detection procedures had very low and nearly indistinguishable Type I error rates when no group ability difference existed and in some cases in which there were differences in group ability.

In simulation research with DIF-free studied items, Type I error for the IRTLR DIF test was well controlled in various situations, as long as most anchor items were actually DIF-free. Type I error rates have been close to the nominal level for 2PL, 3PL, and graded models (Ankenmann et al., 1999; Cohen et al., 1996; Kim & Cohen, 1998; Sweeney, 1997; Wang & Yeh, 2003), and the empirical mean and standard deviation of the likelihood-ratio statistic have been near what they should be for a χ^2 distributed statistic (Ankenmann et al., 1999). In research

by Woods (2009), at all values of the focal group sample size, Type I error was well below the nominal level for the MIMIC approach than for IRTLR. As for the improved Wald test, results indicated that Wald-1 performed very well and is recommended, whereas Type I error was extremely inflated for Wald-2. Performance of IRTLR and Wald-1 was similar (Woods et al., 2013).

Finch (2005) conducted simulations to compare the MIMIC method with IRTLR, and found that when tests contain 50 items following the 2PL model, the MIMIC method was very competitive with IRTLR in terms of better control of Type I error, regardless of the focal group size, differences in mean group abilities, and magnitude of DIF contamination in the matching variable. IRTLR was adversely affected by the magnitude of DIF contamination in the anchor, whereas the MIMIC method showed only a small inflation in the Type I error rates. Thus, if a large number of the items are suspected to exhibit DIF, the MIMIC method is preferable because it's Type I error rates are not as seriously influenced by DIF contamination as those of IRTLR. With varying focal and reference group sizes and scale lengths, Type I error was well-controlled, and estimates of the focal-group mean were quite accurate (Woods, 2009). Type I error rates determine quality of performance and support use of one method over another (depending on experimental condition).

Power

Similar to Type I error rates, power rates vary according to method used and simulation conditions. Power rate is defined as the percentage of DIF items correctly detected out of the amount of replications. It is important to keep in mind the interpretation of theoretical power rates are conditional on the Type I error rates for a given significance level because power rates can artificially increase if the Type I error rates are inflated. In other words, high power can be

due to Type I error inflation (Woods, 2008), so researchers must use caution when interpreting power. In previous simulation research, a cutoff of 0.8 was used to indicate excellent power and power rates between 0.7 and 0.8 were evaluated as moderate (Cohen, 1992). For IRTLR, statistical power to detect uniform and nonuniform DIF increases with increases in sample size, item discrimination, the number of anchors, and the amount of DIF in the data (Ankenmann et al., 1999; Wang & Yeh, 2003; Woods, 2009; Woods et al., 2013). For example, in research by Ankenmann et al. (1999), optimal conditions for maximizing power were larger *a* parameters, larger sample sizes, and identical population means.

The power of a DIF procedure is related directly to sample size. With very small samples of reference and/or focal group members, even items displaying substantial DIF can go undetected. It is typically suggested that larger samples are required for use with IRT methods when two or three parameter models are used (Clauser & Mazor, 1998). For SEM methods, to achieve reasonably powerful and accurate MIMIC results, focal group samples size should be at least around 100. Also, because power is greater when items are more discriminating (Ankenmann et al., 1999), smaller focal group sample size may be acceptable for the highly discriminating items sometimes observed on psychopathology scales (e.g., Rodebaugh, Woods, Thissen, Heimberg, Chambless, & Rapee, 2004).

According to Finch (2005), the power of IRTLR appears to be more influenced by the size of the focal group than does that of the MIMIC method, especially when there is DIF contamination in the anchor items. When considering the ratio of reference to focal group sample size, power was lower when sample sizes were unequal (Kristjansson et al., 2005). It is not surprising that unequal sample sizes can reduce power; large differences in group sample sizes mean that at each level of total score, there are relatively fewer examinees in the focal

group to compare to those in the reference group. Thus, the effective sample size is smaller, and power is lower. In DIF detection research by Hou et al. (2014), power rates increased as the sample size increased, irrespective of other factors. Woods (2009) found that the sample size needed for adequate power and reasonable accurate estimates of most item parameters was smaller for MIMIC models than for IRTLR. However, IRTLR always had greater power to detect nonuniform DIF than MIMIC models, and bias was elevated for some MIMIC model item parameter estimates.

Scale length is unlikely to have much impact on statistical power or item parameter accuracy, but longer scales may produce more accurate estimates of the mean difference (Woods, 2009). In some instances, the MIMIC method has more power than IRTLR when tests are long (Finch, 2005). Wang & Shih (2010) also stated that MIMIC attains a comparable or higher power of DIF detection for long tests (50 items), but not quite as effective for short tests (20 items).

Chang et al. (1996), Tian (1999), and Zwick et al. (1997) have found that most procedures have very high power for uniform DIF when the studied items have moderate and high item discrimination (Kristjansson et al., 2005), and power for uniform DIF improved with increased item discrimination. The reference item parameter values also influenced the power rates. As the reference item parameter values increased, the power rates decreased. DIF magnitude also had a distinctive impact on the power rates. In research by Hou et al. (2014), the larger DIF magnitude corresponded to higher power rates across the uniform DIF types and sample sizes.

Anchor Item Selection

An assumption of DIF analyses is that the metric for different groups is linked using anchor items that are invariant (i.e., equivalently functioning between groups). It is impossible for researchers in practice to discern which items are differentially functioning (DF) and which are invariant (or the opposite of DF). DIF research has long been plagued with anchor item issues, and various approaches have been suggested. The relative efficacy of some of these approaches have been tested (Kopf, Zeileis, & Strobl, 2015; Shih & Wang, 2009; Wang & Yeh, 2003; Woods, 2009), and an easily implemented 2-stage procedure (described in this section) put forth by Lopez Rivas, Stark, and Chernyshenko (2009) and further validated by Meade and Wright (2012) was superior. With this approach, appropriate anchor items can be easily and quickly located, resulting in more efficacious invariance tests, providing optimal power while maintaining nominal Type I error (Meade & Wright, 2012).

IRTLR's most common approach is the constrained baseline model or *all others as anchors* (AOAA) approach (discussed in the IRTLR section). The shortcoming of the AOAA approach is that if any of the scale items are DF, they are included in the anchor item set. Researchers have cautioned that DF items included in the anchor set leads to inflated Type I error (Candell & Drasgow, 1988; Holland & Thayer, 1988; Kim & Cohen, 1995; Lautenschlager, Flaherty, & Park, 1994; Lord, 1980; Navas-Ara & Gomez-Benito, 2002; Park & Lautenschlager, 1990; Wang et al., 2009; Woods, 2009). Furthermore, Type I error increases as the number of DF items in the anchor set increases (Stark et al., 2006; Wang & Yeh, 2003). It is possible to use a single anchor item to help reduce the likelihood of a DF item being used as an anchor. However, if a single anchor item is chosen that is not invariant across groups, Type I error is the result (Johnson, Meade, & DuVernet, 2009; Lopez Rivas et al., 2009). Stark et al. (2006)

showed that when DF items were present, having one invariant anchor item was preferable to the AOAA approach. Adding additional invariant anchor items, however, can increase power (Bolt, 2002; Lopez Rivas et al., 2009; Thissen et al., 1988; Wang & Yeh, 2003; Woods, 2009) and using a single item can result in an underpowered test (Meade & Wright, 2012). Therefore, it is best to have as many invariant anchor items as possible, but Type I error can be increased due to inclusion of a DF item as an anchor (Finch, 2005; Stark et al., 2006). More anchor items means a greater chance one will be DF. It is impossible to be certain which items are DF and which are invariant with non-simulated data, so it appears the optimal approach is one in which power is maximized by having more than one anchor item, and these items exhibit little to no DIF (Meade & Wright, 2012).

The two criteria to evaluate potential methods of identifying anchor items are accuracy and ease of implementing the anchor item selection technique. Tedious and complicated methods will not be regularly used in practice no matter how accurate. For example, an iterative scale purification procedure (Cao et al., 2017; Kim & Cohen, 1995; Tay et al., 2015), despite adequate performance, is burdensome to implement. Anchor items are identified after multiple required runs and output evaluation and reanalysis at each stage. Regardless of the number of iterations performed, there is no guarantee that a stable set of DF items will emerge.

As previously mentioned, for selecting anchor items, Lopez Rivas et al. (2009) recommended using AOAA and then selecting non-DF items with the largest discrimination parameters (*a* parameters) as anchors. A limitation of this approach is not knowing how many anchor items to choose. Stark et al. (2006) found that a single anchor item can work well under some conditions. However, power is known to increase as the number of anchors increases (up to a certain amount) (Bolt, 2002; Lopez Rivas et al., 2009; Thissen et al., 1988; Wang & Yeh,

2003; Woods, 2009). Lopez Rivas et al. (2009) recommended using as many as three anchor items using the same rule (where non-DIF items with the largest *a* parameters are chosen) as selecting a single item. Consistent with other research, (Bolt, 2002; Lopez Rivas et al., 2009; Thissen et al., 1988; Wang & Yeh, 2003), Meade and Wright (2012) found that a larger number of anchor items led to more power which maximized around five anchor items (in a 20 item test; meaning selecting more anchors did not further increase power). They also found that a single anchor item used with 19 test items was insufficient with respect to power. These findings are similar to those of Stark et al. (2006), who also found unacceptable power when using one anchor item. In contrary, Wang et al. (2009) found suitable power with a single anchor item, but their study simulated large DF items. Meade and Wright (2012) recommend using more than one anchor item, especially in cases of mild DF and with 20 items or more in the assessment.

IRTLR is easily implemented and provides a parametric test of DIF. However, selecting the proper anchor items with IRTLR is imperative, and using AOAA is known to result in Type I errors when some items are not invariant (Candell & Drasgow, 1988; Kim & Cohen, 1995; Wang et al., 2009; Woods, 2009; Meade & Wright, 2012). Therefore, it is necessary to identify a set of anchor items that maximize power and minimize Type I errors. According to Meade and Wright (2012), a simple two-stage approach performed as well or better than a labor-intensive iterative scale purification method and much better than the Type I error-proned AOAA approach alone. For the two-stage approach, they recommend conducting IRTLR using the AOAA approach and then select up to five invariant items with the largest *a* parameters to serve as anchor items in an additional final test of invariance using the DIF detection method of your choice. There are five steps to conduct this approach: (a) conduct invariance analyses via the default AOAA approach, (b) examining only non-significant items, identify the five with the

largest *a* parameters, (c) conduct final test of invariance with a free baseline approach using items identified in Step b, (d) evaluate DF significance levels from the DIF method used, and (e) compute DF effect size indices using output from Steps a and c.

Manipulated Factors in Monte Carlo Simulation Research

The performance of IRT and SEM DIF detection methods are influenced by factors such as: sample size, test length, proportion of DIF item contamination, magnitude of DIF, group impact on ability distribution, number of groups, types of tests (Cao et al., 2017; Chun et al., 2016; Finch, 2005; Hou et al., 2014; Kristjansson et al., 2005; Mazor, Clauser, & Hambleton, 1992; Raju et al., 2002; Rogers & Swaminathan, 1993; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang & Shih, 2010; Wang et al., 2009; Woods, 2008; Woods, 2009).

Sample size

Sample size is a critical factor that can affect both the precision of the factor loading estimate and the power in DIF tests (Meade & Bauer, 2007). The power of detecting any true differences across populations is affected by insufficient sample sizes used in previous research (e.g., Schaubroeck & Green, 1989; Vandenberg & Self, 1993). In most of the simulation studies reviewed, sample size was a manipulated parameter (Cao et al., 2017; Chun et al., 2016; Hou et al., 2014; Kristjansson et al., 2005; Navas-Ara & Gomez-Benito, 2002; Oort, 1998; Raju et al., 2002; Stark et al., 2006; Swaminathan and Rogers, 1990; Wang & Shih, 2010; Wang et al., 2009; Woods, 2008; Woods, 2009; Woods et al., 2013). Varying sample sizes often produced different DIF detection outcomes for each method, making it an ideal parameter to manipulate to compare method performance.

Test Length

Many of the previous simulation studies varied test length to see its effect on DIF detection capabilities between methods. Researchers have also encouraged future research to manipulate test length to test DIF detection. For example, Finch (2005) recommended further efforts be made to identify the test length for which MIMIC model performs comparably to other DIF detection methods. To determine factors that affect power, Meade & Lautenschlager (2004) simulated test lengths of six and 12 items. Su & Wang (2005) used 25 and 50 items in their study to compare performances of three polytomous DIF detection methods. Using various DIF detection methods, Swaminathan and Rogers (1990) used test lengths of 40, 60, and 80 for their simulated research. A DIF analysis performed by Zwick et al. (1997) used a set of either 20 or 50 items. Similarly, for MIMIC DIF detection, Wang et al. (2009) used test lengths of 20 and 50 items as well. To explore the DIF detection capabilities of the Wald test, Cao et al. (2017) used test length of 15 and 30. To test the sequential-free baseline MIMIC approach, Chun et al. (2016) used a single test length of 15 items. In a review of the literature, the number of items in a unidimensional set analyzed with IRTLR DIF detection was typically between seven and 38, with a mean over studies (excluding a few outliers) of 20.

Type of DIF

A studied item's discrimination parameter has been shown to impact Type I error and power in simulation research. During investigation of DIF, Chang et al. (1996) and Mazzeo and Chang (1994) found highly inflated Type I error when the studied item's *a* parameter was substantially higher than that of the anchor items. However, there was a positive correlation between the power of uniform DIF and the studied item's *a* parameter (Chang et al., 1996).

Similar findings for DIF detection were found by Roussos and Stout (1996) and Uttaro and Millsap (1994), where *a* parameters had the potential to inflate Type I error.

In other simulation research, Kristjansson et al. (2005) varied type of DIF in three different ways (null, uniform, nonuniform) to see how methods performed detecting different types of DIF. For their DIF detection method performance comparison, Chun et al. (2016) simulated type of DIF in a similar way: none, threshold, loading. Woods (2008, 2009) also simulated the presence or absence of DIF in the data in a similar fashion. Stark et al. (2006) simulated DIF type when it was present: (a) DIF on thresholds only, (b) DIF on loadings only, or (c) DIF on both loadings and thresholds. Wang et al. (2009) simulated DIF pattern, with DIF being one-sided (uniform or nonuniform) or balanced (both uniform and nonuniform). Hou et al. (2014) manipulated DIF type in their research to detect DIF as well. It is unlikely in practice that all items favor the same group (Woods, 2008), which is why it is important to make these manipulations to simulate real testing conditions.

DIF Contamination

Previous simulation studies manipulated the proportion of DIF contamination in the test items. In a simulation study focused on selecting appropriate anchors for DIF detection, Woods (2009) simulated 0%, 20%, 50%, or 80% DF test items. To study MIMIC DIF detection capabilities, Wang et al. (2009) simulated percentage of DIF items in the test (i.e., 0%, 10%, 20%, 30% and 40%). In a Monte Carlo simulation to explore the iterative Wald test for DIF detection, Cao et al. (2017) manipulated percentage of DIF items (i.e., 20%, 40%). In a simulation study by Swaminathan and Rogers (1990), 20% items with DIF were used. Woods (2008) also manipulated the presence and absence of DIF in the data. Five of the 24 items (20.8%) analyzed had nonuniform DIF and 10 out of 24 (41.7%) items had uniform DIF.

DIF Magnitude

The magnitude of DIF added or subtracted from baseline parameters was another common manipulated condition in previous simulation research. For example, Stark et al. (2006) simulated different amounts of DIF: (a) no DIF, where reference and focal group loading (discrimination) and threshold (difficulty) parameters were set equal for all items; (b) small DIF, where for focal group, loadings were decreased by 0.15 and item thresholds increased by 0.25, with respect to reference values; or (c) large DIF, where focal group loadings decreased by 0.4 and thresholds increased by 0.5. In research by Woods et al. (2013) comparing DIF detection performance of the Wald test and IRTLR, small, medium, and large amounts of DIF were simulated at levels of 0.3, 0.5, and 0.7, respectively. Similar to Woods et al. (2013), to simulate uniform and nonuniform DIF effects, Cao et al. (2017) added or subtracted a constant of 0.3, 0.5, and 0.7 for different effect size conditions. Performance of four methods of DIF detection were compared by Kristjansson et al. (2005), where uniform DIF was simulated by increasing *b* parameters by 0.25. For nonuniform DIF, varying amounts were added according to the studied item discrimination. The *a* parameter was 1.0 higher for the reference group when the studied item discrimination was 0.8, 1.3 higher when studied item discrimination was 1.2, and 1.6 higher when studied item discrimination was 1.6. According to the authors, these values were chosen to make DIF magnitude in the uniform and nonuniform conditions approximately equivalent. This approach was also used by Swaminathan and Rogers (1990) to produce uniform and nonuniform DIF in simulated data at equivalent levels. Lee et al. (2017) manipulated focal group item parameters to simulate uniform and nonuniform DIF for method performance comparison. For uniform DIF, the studied item's *b* parameter was increased by 0.25 to represent a low level of DIF magnitude, and 0.5 to represent a medium level. For nonuniform DIF, 0.3 and 0.6 were

used to represent low and medium levels of DIF magnitude, respectively. Oshima, Raju, & Flowers (1997) and Suh and Cho (2014) used these level differences in parameters between reference and focal groups in IRT DIF studies as well. To simulate nonuniform DIF, Chun et al. (2016) decreased focal group loadings by 0.15; for uniform DIF, focal group thresholds were increased by 0.25.

Recommendations from Prior DIF Research

There have been many simulations studies on technical issues of DIF, and as research progresses, many methodological problems appear (Zumbo, 2007). The efficiency and accuracy of DIF methods using IRT and SEM have been reviewed in many of these simulation studies (Cao et al., 2017; Chun et al., 2016; Finch, 2005; Hou et al., 2014; Kristjansson et al., 2005; Navas-Ara & Gomez-Benito, 2002; Oort, 1998; Raju et al., 2002; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang & Shih, 2010; Wang et al., 2009; Woods, 2008; Woods, 2009; Woods et al., 2013). Both SEM and IRT models provide interesting ways of representing data in the social and behavioral sciences, and simulation studies can address questions left unanswered by researchers in these fields. Based on findings from these simulation studies, researchers have made many recommendations for future researchers that have guided the formulation of the current research questions and the study design to address them. Also, these researchers have encouraged future investigators to search for ways of deciding which DIF model is best for which research purpose (e.g., Reise et al., 1993). The outcome of such work would be a more coherent framework for the use of current, state-of-the-art methods of psychometric analysis. Given increasing interest in the MIMIC method for DIF detection and the relative lack of thorough simulations assessing its effectiveness, Wang et al. (2009)

encouraged researchers to verify what relative advantages and disadvantages the MIMIC method has compared to other popular DIF detection methods. Therefore, important questions such as which method works best depending on simulation conditions, and the comparability of results should be examined more comprehensively in future investigations.

After comparing IRTLR and MIMIC modeling for DIF detection, Woods et al. (2009) recommended future researchers use additional samples because the pattern of DIF and the parameter estimates observed in some samples may not be the same for other samples. For example, based on Woods et al.'s (2009) recommendations, researchers need to assess DIF with larger focal group sample sizes, either with MIMIC or IRTLR. The accuracy of the MIMIC approach to DIF testing is incompletely verified, and little is known about the sample size requirements (Woods, 2009), so manipulating sample size would help validate the approach. Kristjansson et al. (2005) similarly stated that the evaluation of the effect of different sample sizes is needed to understand how DIF methods might perform in applied testing situations.

Also, the capabilities of MIMIC-interaction models that detect nonuniform DIF can be further explored and validated with additional types of simulated data (Woods & Grimm, 2011). This follows the recommendations of Zumbo (2007), who stated that nonuniform DIF is a problem that has plagued researchers for decades and should be focused on in simulation research.

At first, MIMIC model implementation was straightforward only when nonuniform DIF testing was ignored. Now, however, MIMIC-interaction models (Woods & Grimm, 2011), which test for nonuniform DIF, have become more accurate (Lee et al., 2017). Woods et al. (2013) claimed that once this occurred, a comparison between MIMIC method and Wald test needed to be accomplished. The Wald test has yet to be compared with other DIF detection methods, so

future researchers were encouraged to examine the DIF detection performance of the Wald test in comparison to other competitive DIF detection methods (i.e., MIMIC and IRTLR) (Woods et al., 2013; Cao et al., 2017; Tay et al., 2015). The simulation study design for the current research is based on recommendations from these studies.

Research Questions

Based on recommendations and manipulation factors suggested from previous research, two overarching research questions were investigated. The first research question investigated each method's rate of incorrect DIF identification, measured by Type I error, by asking the following:

1. *With regard to rate of incorrect DIF identification (Type I error), which method among MIMIC, IRTLR, and the Wald test will perform best in each scenario of simulation?*

To continue performance comparison, a second research question investigated each method's ability to detect DIF, measured by power. The second research question asked the following:

2. *With regard to accuracy to detect DIF (power), which method among MIMIC, IRTLR, and the Wald test will perform best in each scenario of simulation?*

CHAPTER 3

METHOD

Simulation Design

Data Generation

Dichotomously scored items are a typical type of non-continuous data that have been frequently used in assessments and questionnaires in the social sciences. Previous research has been performed using dichotomous data to compare DIF detection capabilities under both IRT and SEM frameworks (Cao et al., 2017; Chun et al., 2016; Kim et al., 2012; Raju et al., 2002; Stark et al., 2006; Woods & Grimm, 2011). Therefore, dichotomous data was used to compare the DIF detection performance of the MIMIC method and the IRTLR and Wald tests. To construct the current simulation, published applications of the methods of interest were used as guides. Multiple algorithms were written in an R script, version 3.4.2 (R Core Team, 2017) to automate the DIF testing for each method. First, a Monte Carlo simulation was performed to simulate data sets. Item parameters for the SAT verbal items (e.g., Donoghue and Allen, 1993; Finch, 2005) were used for simulating dichotomous item responses, and the latent trait for examinees, θ, was drawn from a normal distribution with a mean and variance of 0 and 1, respectively. The 3PL model (see formula 1) was used to generate the data as in previous successful DIF studies (Finch, 2005; Lopez Rivas et al., 2009; Shih, 2009; Wang et al. 2009).

Data sets were manipulated for each simulation and generated to have the varying experimental conditions (i.e., manipulated factors) listed below. With a total of 52 experimental conditions, 100 replications were performed for each experimental condition, as in previous DIF

detection research (Chun et al., 2016; Lee et al., 2017; Lopez Rivas et al., 2009; Shih, 2009; Wang & Shih, 2010; Wang et al., 2009; Woods, 2009; Woods and Grimm, 2011).

Table 1
Item parameter values used to generate dichotomous item responses

Item	*a*	*b*	*c*
1	1.10	-0.70	0.20
2	0.70	-0.60	0.20
3	0.90	-0.40	0.20
4	1.40	0.10	0.20
5	0.90	0.90	0.16
6	1.20	0.70	0.12
7	0.90	0.30	0.20
8	0.40	0.80	0.20
9	1.60	1.10	0.06
10	2.00	1.10	0.05
11	0.90	-1.50	0.20
12	1.40	-0.40	0.20
13	1.60	-0.10	0.16
14	1.20	0.50	0.20
15	1.20	1.40	0.11
16	1.80	1.40	0.12
17	2.00	1.60	0.16
18	1.00	1.60	0.13
19	1.50	1.70	0.09
20	1.20	1.60	0.09
21	0.70	-0.50	0.20
22	1.20	-0.30	0.20
23	0.90	0.20	0.20
24	0.70	-0.40	0.20
25	0.60	0.20	0.20

Independent Variables in Simulation

Simulated experimental conditions were created by manipulating four independent variables before running each simulation: Test length was set to 25 items. Similar test lengths were used in previous research to compare DIF detection capabilities (Cao et al., 2017; Chun et

al., 2016; Bolt, 2002; Finch, 2005; Hong et al., 2008; Stark et al., 2006; Su & Wang, 2005; Zwick et al., 1997).

1. Reference (R)/Focal (F) group sample size:

    (a) R/1000, F/1,000 (b) R/500, F/500 (c) R/1,500, F/500 (d) R/750, F/250

    The sample size ratio combinations were chosen because they were similar to previous simulation studies (Cao et al., 2017; Lee et al., 2017; Woods et al., 2013).

2. Type of DIF: Null, Uniform, Nonuniform

    Three DIF types were simulated in the studied item (null, uniform, and nonuniform). For the null DIF test, item parameters were not manipulated before simulation. For uniform DIF testing, focal *a* parameters remained the same, but focal *b* parameters were increased before simulation. For nonuniform DIF testing, focal *b* parameters remained the same, but focal *a* parameters were decreased before simulation. The number of items manipulated and the amount of increase or decrease was dependent on the proportion of simulated DIF items and the DIF magnitude being tested. Reference parameters remained the same for all tests types.

3. Proportion of simulated DIF items: 0%, 20%, 40%

    The proportion of simulated DIF items was manipulated by having none of the items simulated with DIF for the null condition, simulating DIF in items 21-25 for the 20% condition, and simulating DIF in items 16-25 for the 40% condition.

4. DIF magnitude: ±0.3, ±0.5, ±0.7

    The *a* and *b* parameters were unaltered for the reference and focal groups in the null-DIF condition. In the uniform DIF condition, *a* parameters for both groups were unaltered, but *b* parameters for the focal group were increased by either 0.3, 0.5, or 0.7,

depending upon whether small, medium, or large DIF magnitude was being tested. That way, the focal group would be less likely than the reference group to achieve the higher score. A number of researchers have used this level of uniform DIF (e.g., Ankenmann et al., 1999; Chang et al., 1996; Lee et al., 2017; Kristjansson et al., 2005; Spray & Miller, 1994; Tian, 1999; Zwick et al., 1997). In the nonuniform DIF condition, *b* parameters were unaltered for both groups, but the *a* parameter for the focal group was decreased by 0.3, 0.5, or 0.7, depending, again, on the magnitude of DIF being tested.

Table 2
Combination of independent variables for each simulation

Simulation	Reference	Focal	DIF %	DIF Type	DIF Level
1	750	250	0	Null	No DIF
2			20	Uniform	Small (+0.3)
3					Med. (+0.5)
4					Large (+0.7)
5				NonUniform	Small (-0.3)
6					Med. (-0.5)
7					Large (-0.7)
8			40	Uniform	Small (+0.3)
9					Med. (+0.5)
10					Large (+0.7)
11				NonUniform	Small (-0.3)
12					Med. (-0.5)
13					Large (-0.7)
14	500	500	0	Null	No DIF
15			20	Uniform	Small (+0.3)
16					Med. (+0.5)
17					Large (+0.7)
18				NonUniform	Small (-0.3)
19					Med. (-0.5)
20					Large (-0.7)
21			40	Uniform	Small (+0.3)
22					Med. (+0.5)
23					Large (+0.7)
24				NonUniform	Small (-0.3)
25					Med. (-0.5)
26					Large (-0.7)

Table 2
Combination of independent variables for each simulation

Simulation	Reference	Focal	DIF %	DIF Type	DIF Level
27			0	Null	No DIF
28					Small (+0.3)
29				Uniform	Med. (+0.5)
30			20		Large (+0.7)
31					Small (-0.3)
32				NonUniform	Med. (-0.5)
33	1500	500			Large (-0.7)
34					Small (+0.3)
35				Uniform	Med. (+0.5)
36			40		Large (+0.7)
37					Small (-0.3)
38				NonUniform	Med. (-0.5)
39					Large (-0.7)
40			0	Null	No DIF
41					Small (+0.3)
42				Uniform	Med. (+0.5)
43			20		Large (+0.7)
44					Small (-0.3)
45				NonUniform	Med. (-0.5)
46	1000	1000			Large (-0.7)
47					Small (+0.3)
48				Uniform	Med. (+0.5)
49			40		Large (+0.7)
50					Small (-0.3)
51				NonUniform	Med. (-0.5)
52					Large (-0.7)

Evaluation Criteria

The evaluation criteria to determine quality of performance in each simulation condition was Type I error and power. To compute Type I error, the number of non-DIF items incorrectly identified as DIF items was divided by the total number of non-DIF items in the scale. To calculate statistical power, the number of DIF items correctly detected by each method was divided by the total number of DIF items in the scale. Indications of better performance of a DIF detection method was signified by Type I error rates well-controlled at or below the nominal

Type I error rate of 0.05, and higher statistical power. All reviewed simulations studies that used Type I error rate and power to determine quality of performance of a DIF detection method were computed in a similar fashion (Cao et al., 2017; Chun et al., 2016; Finch, 2005; Hou et al., 2014; Kim et al., 2012; Kristjansson et al., 2005; Stark et al., 2006; Swaminathan & Rogers, 1990; Wang et al., 2009; Woods, 2008; Woods, 2009; Woods, 2009b; Woods & Grimm, 2011; Woods et al., 2013).

DIF Analyses

Software for DIF Analysis

For the current research, R version 3.4.2 (R Core Team, 2017) was used to perform the Monte Carlo simulation to create data as well as select anchors, detect DIF, and calculate Type I error and power for each of the analysis methods. For the IRTLR and Wald DIF tests, the R packages used for analysis were *mirt* (Chalmers, 2012), *psych* (Revelle, 2017), *devtools* (Wickham & Chang, 2016) and *lessR* (Gerbing, 2012). For the MIMIC DIF test, the *MplusAutomation* R package (Hallquist & Wiley, 2018) was used in conjunction with Mplus Version 8 (L.K. Muthen & Muthen, 1998-2018).

MIMIC Method

To use the MIMIC method for DIF testing, two samples were combined to one data set, and a group variable was included as an exogenous variable for the group test. A baseline model was created that did not allow any association between the group variable and errors for items. This model was compared with a model that allows for association between group and error terms for the items. Because the factor loadings are assumed to be equivalent across groups in

the MIMIC model, the association between error terms and the group variable indicates a lack of measurement equivalence (i.e., DIF) between the two samples.

At least one DIF-free item is needed to define the factor on which the groups are matched. Therefore, preliminary analyses were performed to select a subset of DIF-free items to define the factor in subsequent analyses. Every item was tested for DIF with all other items presumed DIF-free by conducting a constrained baseline model test using IRTLR and the anchor item selection method from Meade and Wright (2012). Following their method, five items with the highest discrimination parameters were chosen as anchor items for the final test for DIF. Note that for every simulation, the same anchors were used between methods to simplify the analysis and maximize accuracy and fairness.

Items not assigned to the DIF-free subset of anchors (i.e., studied items) were tested individually for DIF. To test a studied item for DIF, the full model was compared to the constrained model. In both the full and constrained models, all of the original items from the scale were used, and anchor items were not regressed on the group variable. In the full model, all studied items were permitted to have DIF (i.e., all studied items were regressed on the group variable). In the constrained model, invariance (measurement equivalence) was presumed for the studied item (i.e., the studied item was not regressed on the group variable). A significant difference between these models indicates that fit significantly declines if the studied item is assumed DIF-free (i.e., the studied item is DF).

In sum, following the approaches of Thissen and Steinberg (1988), Rivas et al. (2008), Woods & Grimm (2011; for nonuniform DIF detection using MIMIC) and Chun et al. (2016), a free baseline approach was used to detect DIF. The DIF analysis with MIMIC method was performed in two steps: 1) conduct constrained baseline tests to identify items that appear DIF-

free using the anchor item selection from Meade and Wright (2012), and 2) choose the five most discriminating DIF-free items as the anchors for subsequent free baseline DIF tests of the other items in the scale.

IRTLR Method

In IRTLR DIF detection, several two-group IRT models were statistically compared with varying constraints. This is done separately for every studied item (i.e., item to be tested for DIF). For preliminary analysis and as described above, items were chosen as anchors using IRTLR and the method from Meade and Wright (2012). These anchors were assumed to be DIF-free and were used to set a common scale for θ in the final test for DIF. In other words, the parameters of these anchor items were constrained equal between groups, whereas studied items were evaluated for DIF. The θ distribution was assumed normal for both groups with the mean and standard deviation fixed to 0 and 1, respectively, for the reference group and estimated for the focal group simultaneously with the item parameters.

Three DIF tests were of interest for each studied item. First was an omnibus test for DIF in *a* and *b* simultaneously, carried out for a single item at a time. This test was omnibus because it could be significant if there is DIF in *a*, DIF in *b*, or both.

To carry out this test, a model with both parameters for the studied item constrained equal between groups was compared to a model with both parameters for the studied item permitted to vary between groups. In both models, both parameters for all anchor items were constrained equal between groups. The χ^2 distributed test statistic is -2 times the difference between optimized log likelihoods, with *df* equal to the difference in the number of free parameters. With three parameters per item, $df = 3$. Statistical significance indicates the presence of DIF.

Following a significant omnibus test, more specific tests of uniform and nonuniform DIF were carried out. LR test of nonuniform DIF is a test of DIF with respect to the *a* parameter. The models being compared differ in whether the *a* parameter for the studied item was permitted to vary between groups. In both models, the *b* parameter is free to vary between groups for the studied item, and both *a* and *b* are constrained equal between groups for all anchor items. The *b* parameter was free to vary between groups because if the item has DIF in *a*, it may be measuring a different latent variable in one group versus another, in which case there would be no justification for assuming the *b*s are equal. Statistical significance indicates the presence of nonuniform DIF.

For the test of uniform DIF, the models being compared differ in whether *b* for the studied item was permitted to vary between groups. The test of uniform DIF is a test of DIF with respect to *b*, conditional on the absence of DIF in *a*. Therefore, it is reasonable only when the test for nonuniform DIF is non-significant. In both models, *a* was constrained to be group equivalent for the studied item, and both *a* and *b* were constrained to be group equivalent for all anchor items. Statistical significance indicates the presence of uniform DIF.

Wald Test

The Wald-1 method from Cai et al. (2011) requires user-specified anchor items (Woods et al., 2103), so the same method from Meade and Wright (2012) used in both MIMIC and IRTLR was used to select anchor items as matching variables between groups for the Wald test. This strategy of using Wald test could be considered a modified version of the iterative Wald approach from Cao et al. (2017), where IRTLR is used to identify anchor items rather than Wald-2, and then the Wald-1 test is conducted. To identify the scale for the Wald test, a single model is fitted, with reference group mean and standard deviation fixed to 0 and 1. Using parameter

estimation, the mean and standard deviation were simultaneously estimated. Item parameters either allowed studied items to freely vary between groups, or constrained anchor items to be equal between groups. Every studied item produced a Wald statistic.

CHAPTER 4

RESULTS

Results for the DIF experiments are shown in Table 3. In general, IRTLR, Wald, and MIMIC methods suffered from low power (<0.6) for all small magnitude DIF tests (where 0.3 was either added to *b* parameters or subtracted from *a* parameters for simulating uniform and nonuniform DIF, respectively) and many medium magnitude DIF tests (where 0.5 was either added to *b* parameters or subtracted from *a* parameters for simulating uniform and nonuniform DIF, respectively) as well. This was regardless of sample size, type of DIF test, or proportion of items simulated to be DF.

Table 3

Final Results

Experiment	Reference	Focal	DIF %	DIF Type	DIF Level	IRTLR		Wald		MIMIC	
						Type I	Power	Type I	Power	Type I	Power
1	750	250	0	Null	No DIF	0.040	0.000	0.055	0.000	0.089	0.000
2			20	Uniform	Small (+0.3)	0.042	0.162	0.066	0.224	0.063	0.244
3					Med. (+0.5)	0.049	0.466	0.071	0.568	0.071	0.626
4					Large (+0.7)	0.027	0.74	0.053	0.844	0.093	0.868
5				NonUniform	Small (-0.3)	0.034	0.104	0.059	0.188	0.048	0.218
6					Med. (-0.5)	0.034	0.294	0.05	0.402	0.051	0.528
7					Large (-0.7)	0.039	0.624	0.06	0.722	0.051	0.805
8			40	Uniform	Small (+0.3)	0.054	0.160	0.08	0.196	0.084	0.152
9					Med. (+0.5)	0.054	0.329	0.078	0.416	0.180	0.321
10					Large (+0.7)	0.071	0.564	0.078	0.664	0.245	0.505
11				NonUniform	Small (-0.3)	0.059	0.115	0.081	0.190	0.067	0.154
12					Med. (-0.5)	0.044	0.233	0.069	0.356	0.068	0.338
13					Large (-0.7)	0.046	0.447	0.060	0.608	0.076	0.592
14	500	500	0	Null	No DIF	0.035	0.000	0.073	0.000	0.093	0.000
15			20	Uniform	Small (+0.3)	0.039	0.202	0.081	0.274	0.070	0.38
16					Med. (+0.5)	0.030	0.538	0.069	0.630	0.075	0.736
17					Large (+0.7)	0.039	0.902	0.075	0.940	0.104	0.966
18				NonUniform	Small (-0.3)	0.039	0.122	0.077	0.216	0.049	0.238
19					Med. (-0.5)	0.046	0.458	0.079	0.566	0.051	0.646
20					Large (-0.7)	0.039	0.730	0.065	0.818	0.059	0.882
21			40	Uniform	Small (+0.3)	0.039	0.158	0.077	0.219	0.092	0.156
22					Med. (+0.5)	0.035	0.403	0.084	0.520	0.190	0.397
23					Large (+0.7)	0.058	0.703	0.104	0.749	0.330	0.624
24				NonUniform	Small (-0.3)	0.048	0.115	0.095	0.192	0.082	0.169
25					Med. (-0.5)	0.038	0.305	0.073	0.454	0.092	0.44
26					Large (-0.7)	0.057	0.552	0.084	0.667	0.097	0.684

Table 3

Final Results

Experiment	Reference	Focal	DIF %	DIF Type	DIF Level	IRTLR		Wald		MIMIC	
						Type I	Power	Type I	Power	Type I	Power
27	1500	500	0	Null	No DIF	0.029	0.000	0.061	0.000	0.098	0.000
28			20	Uniform	Small (+0.3)	0.027	0.312	0.060	0.432	0.074	0.476
29					Med. (+0.5)	0.045	0.720	0.075	0.906	0.091	0.920
30					Large (+0.7)	0.051	0.952	0.065	0.988	0.123	0.994
31				NonUniform	Small (-0.3)	0.047	0.184	0.061	0.296	0.055	0.352
32					Med. (-0.5)	0.035	0.510	0.071	0.722	0.067	0.798
33					Large (-0.7)	0.037	0.862	0.061	0.932	0.063	0.974
34			40	Uniform	Small (+0.3)	0.056	0.215	0.079	0.336	0.134	0.244
35					Med. (+0.5)	0.052	0.525	0.090	0.714	0.283	0.518
36					Large (+0.7)	0.053	0.805	0.084	0.894	0.444	0.736
37				NonUniform	Small (-0.3)	0.042	0.132	0.078	0.277	0.074	0.265
38					Med. (-0.5)	0.048	0.409	0.062	0.603	0.091	0.569
39					Large (-0.7)	0.055	0.649	0.083	0.819	0.119	0.775
40	1000	1000	0	Null	No DIF	0.057	0.000	0.083	0.000	0.105	0.000
41			20	Uniform	Small (+0.3)	0.047	0.382	0.077	0.510	0.088	0.602
42					Med. (+0.5)	0.044	0.836	0.069	0.924	0.113	0.958
43					Large (+0.7)	0.034	0.992	0.073	0.996	0.142	1.000
44				NonUniform	Small (-0.3)	0.044	0.218	0.078	0.400	0.061	0.458
45					Med. (-0.5)	0.053	0.658	0.065	0.792	0.054	0.866
46					Large (-0.7)	0.044	0.934	0.063	0.952	0.060	0.992
47			40	Uniform	Small (+0.3)	0.048	0.289	0.070	0.429	0.138	0.322
48					Med. (+0.5)	0.047	0.648	0.078	0.818	0.353	0.639
49					Large (+0.7)	0.058	0.884	0.120	0.941	0.590	0.812
50				NonUniform	Small (-0.3)	0.070	0.228	0.079	0.367	0.088	0.332
51					Med. (-0.5)	0.052	0.531	0.074	0.711	0.121	0.685
52					Large (-0.7)	0.050	0.799	0.069	0.868	0.126	0.867

In Figure 6, when testing for nonuniform DIF with R/750, F/250 sample size, proportion of DIF items at 20%, and a medium magnitude of DIF, power rates for IRTLR, Wald test, and MIMIC method were 0.29, 0.40, and 0.53, respectively. However, power rate did increase for all methods with sample size increase. For example, when sample size increased to R/1500, F/500, power rates for each method rose to 0.51, 0.72, and 0.80, respectively. Similar increases in power were seen when sample size increased, regardless of the other conditions.

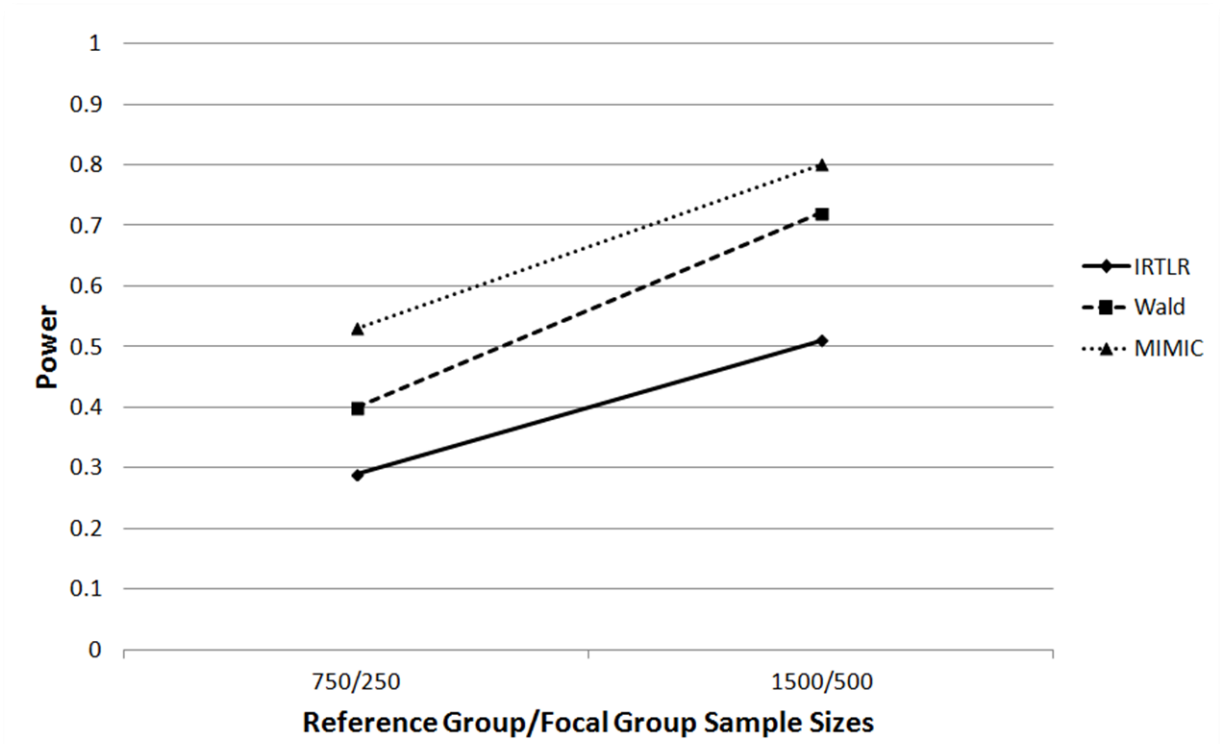


Figure 6. **20% Nonuniform Medium Magnitude DIF**

When comparing the proportion of simulated DIF items at 20% and 40%, the power rate was greater when the proportion of DIF items was at 20%, regardless of the detection method used. In Figure 7, when testing for uniform DIF with sample size R/500, F/500 and large DIF

magnitude, when proportion of DIF contamination rose from 20% to 40%, the power rate dropped from 0.90 to 0.70 for IRTLR test, 0.94 to 0.75 for the Wald test, and 0.97 to 0.62 for the MIMIC method.

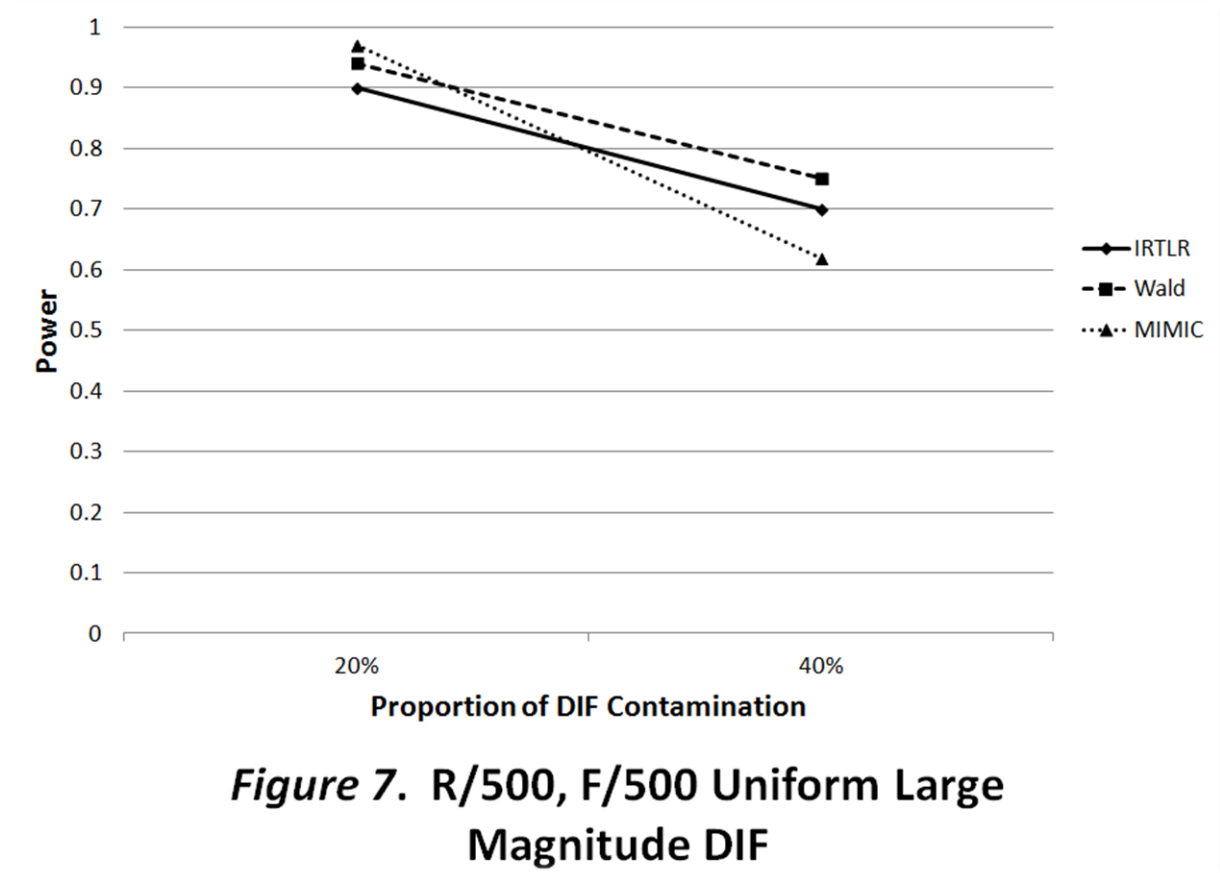


Figure 7. **R/500, F/500 Uniform Large Magnitude DIF**

Also, regardless of the method used, the larger the DIF magnitude, the greater the difference in power rates between 20% and 40% DIF simulations. In Figure 8, with R/750, F/250 sample size and a small magnitude of uniform DIF, the MIMIC method had very little difference in power rate between 20% and 40% DIF contamination conditions, with rates of 0.24 and 0.15, respectively. However, when magnitude of DIF increased to a medium level, power rate for the 20% condition was 0.63, whereas the 40% condition was 0.32. The difference

continued to increase with large DIF magnitude, where power rate for the 20% condition was 0.87 and 0.51 for the 40% condition.

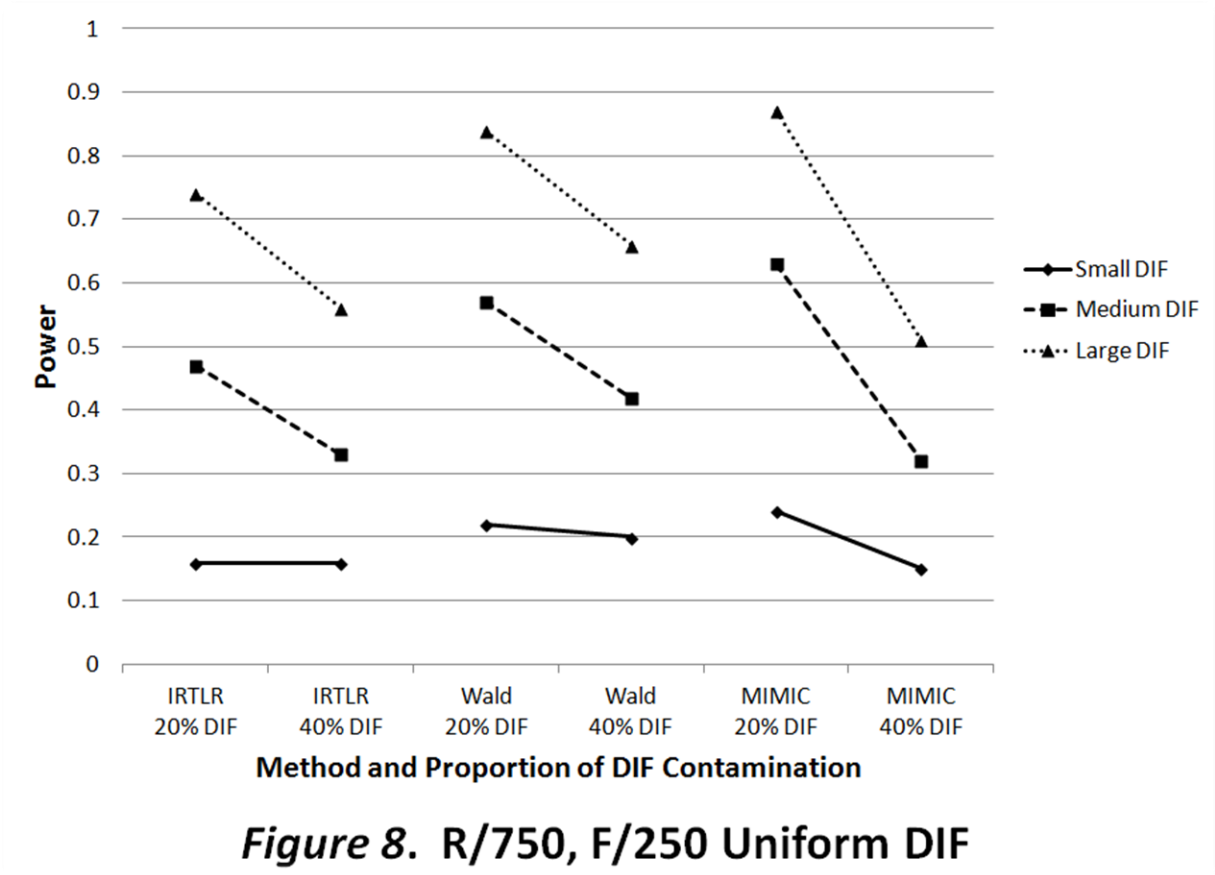


***Figure 8*. R/750, F/250 Uniform DIF**

For IRTLR and Wald tests, uniform DIF detection had a higher power rate than nonuniform DIF detection, regardless of sample size, DIF magnitude, or proportion of DIF. In Figure 9, when using IRTLR to test for a medium level of uniform DIF with R/1500, F/500 sample size and proportion of DIF items at 20%, power rate was 0.72. Power rate of the nonuniform test with the same conditions was 0.51. In the previous two simulation conditions, the Wald test and MIMIC method had less difference in power rates between uniform and nonuniform tests, with power rates dropping from 0.91 to 0.72, and 0.92 to 0.80, respectively.

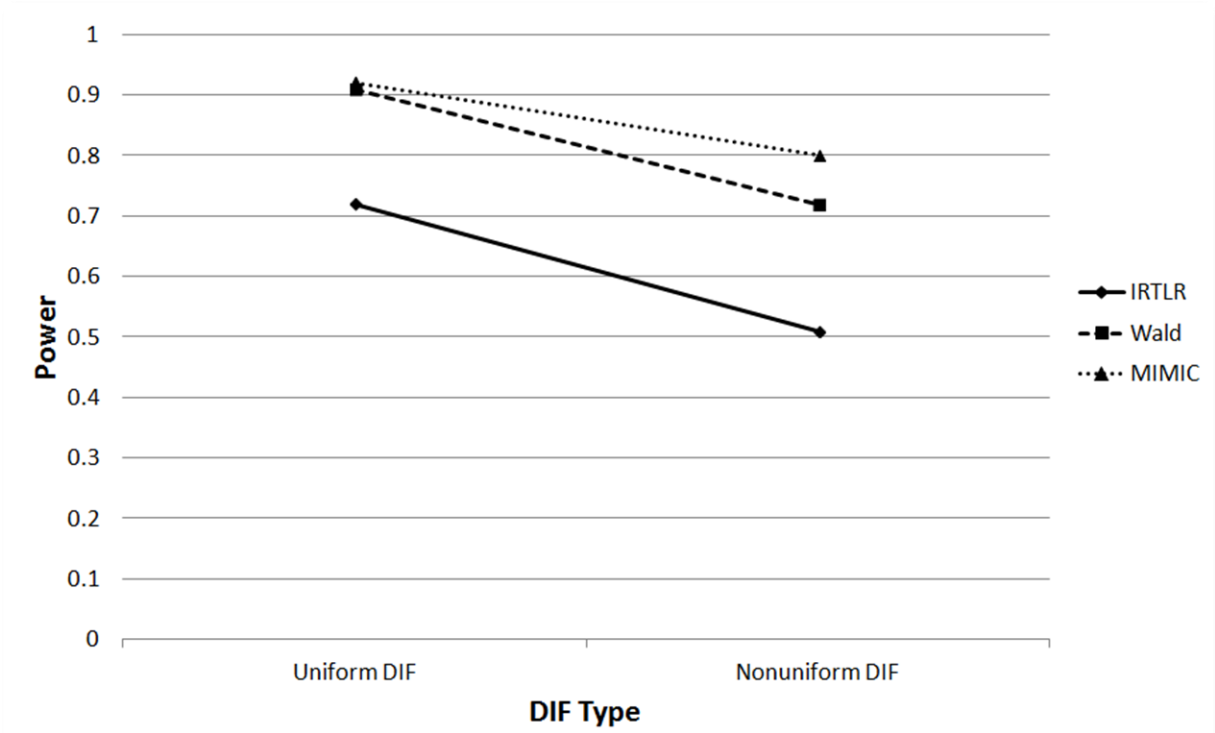


***Figure 9*. R/1500, F/500 20% Medium Magnitude DIF**

Contrary to IRTLR and Wald tests, with MIMIC method, nonuniform DIF detection had more power than uniform DIF detection but only when the proportion of simulated DIF items was 40% in certain sample sizes and DIF magnitudes. Also, most of the differences between uniform and nonuniform DIF power rates in these conditions were less than 0.1. Regardless of sample size, when the proportion of DIF items was 20% in both uniform and nonuniform tests, the MIMIC method had the highest power rate in comparison to IRTLR and Wald test. For example, in Figure 9, the MIMIC method had a power rate of 0.80 when testing for nonuniform DIF with R/1500, F/500 sample size, proportion of DIF at 20%, and a medium magnitude of DIF, whereas Wald and IRTLR tests had power rates of 0.72 and 0.51, respectively. However,

Type I error rate was above the nominal 0.05 for the MIMIC method for most of the simulated DIF experiments (including the previously mentioned example, where Type I error rate was 0.07), regardless of sample size, type of DIF test, or proportion of DIF contamination. In some cases, Type I error rate was severely inflated for the MIMIC method, especially when the proportion of DIF items was 40%. For example, Figure 10 depicts a uniform DIF test with R/1000, F/1000 sample size, proportion of DIF items at 40%, and a large magnitude of DIF, although power level was 0.81 for the MIMIC method, Type I error rate was 0.59.

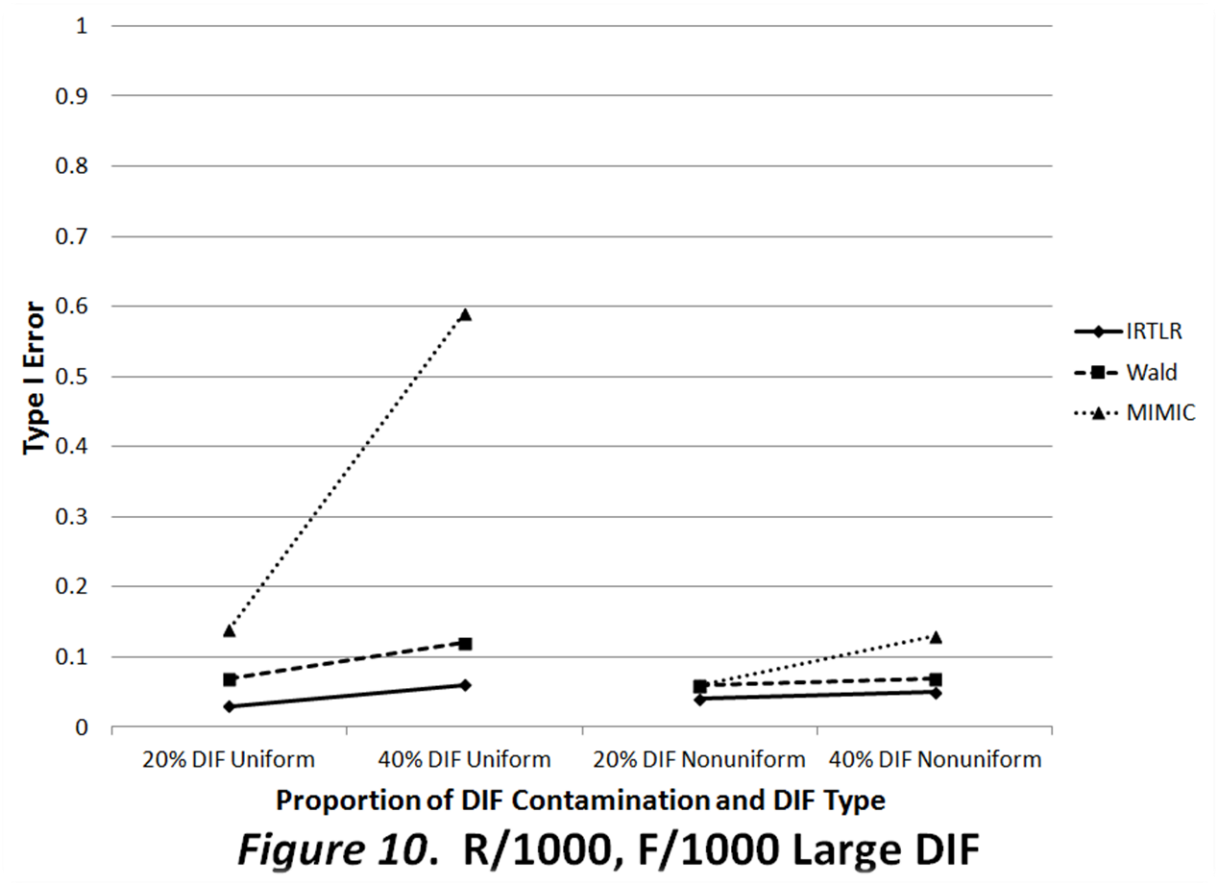


Figure 10. R/1000, F/1000 Large DIF

However, when the proportion of items contaminated with DIF was at 20% for both the uniform and nonuniform tests, MIMIC Type I error rates were close to the acceptable 0.05

threshold. For example, in a nonuniform DIF test with R/750, F/250 sample size, proportion of DIF items at 20%, and a large magnitude of DIF, MIMIC Type I error rate was 0.05. Although MIMIC method had decent Type I error rates for some of the conditions, it never performed better than IRTLR and rarely performed better than Wald test concerning Type I error rate, having the highest Type I error rates in most of the experiments. In Figure 10, the Type I error rate for the MIMIC method was 0.59, whereas the Type I error rates for Wald and IRTLR tests were only 0.12 and 0.06, respectively. In most instances, however, Type I error rate for MIMIC method was close to that of the other methods. IRTLR never had the highest power in any simulation in comparison to Wald test and MIMIC method, but this is not to say IRTLR did not achieve excellent power (>0.8; Cohen, 1992); in many of the large magnitude (where 0.7 was either added to *b* parameters or subtracted from *a* parameters for simulating uniform and nonuniform DIF, respectively) simulation conditions, and some medium magnitude conditions, IRTLR had excellent power. This was especially the case when sample sizes were larger (e.g., R/1500, F/500; R/1000, F/1000), and it did not matter whether it was a uniform or nonuniform DIF test. For example, with R/1000, F/1000 sample size, 20% proportion of DIF items, and large DIF magnitude, IRTLR had a power rate of 0.99 in the uniform DIF test and 0.93 in the nonuniform test. Also, IRTLR was the only method to maintain a Type I error rate at or below 0.05 in most simulations conditions, regardless of sample size, type of DIF test, or proportion of DIF contamination. Therefore, IRTLR had the lowest Type I error rate in comparison to Wald test and MIMIC method throughout every simulation. Differences between Type I error rates for IRTLR and Wald tests were not as large as those between IRTLR and MIMIC method. Looking again at Figure 10, the Type I error rate for IRTLR was 0.06; Wald test had a difference of only 0.06 in Type I error rate in comparison to IRTLR test, whereas the difference between IRTLR

test and MIMIC method was 0.53. Wald test had the highest power in comparison to IRTLR test and MIMIC method when the proportion of DIF was 40%. This was regardless of sample size, DIF magnitude, or DIF type. For example, when the proportion of DIF items was at 40% in a uniform DIF test with R/1500, F/500 sample size and a medium magnitude of DIF, Wald test had a power rate of 0.71, whereas IRTLR and MIMIC method had power rates of 0.53 and 0.52, in that order. Although the power rate for the Wald test was better than its competitors when the proportion of DIF was 40%, the power would be considered low to moderate (<0.8) in many of the small and medium DIF magnitude conditions. In Figure 8, when the DIF magnitude was small, Wald test had a power rate of only 0.22. Also, the Wald test had a Type I error rate greater than 0.05 in 23 out of 25 experiments, although only two of these experiments had Type I error that exceeded 0.1.

To support previous specific examples, results by Type I error rate and power for the large, medium, and small magnitude uniform and nonuniform tests with 20% and 40% proportion of DIF are shown in Figures 11-22.

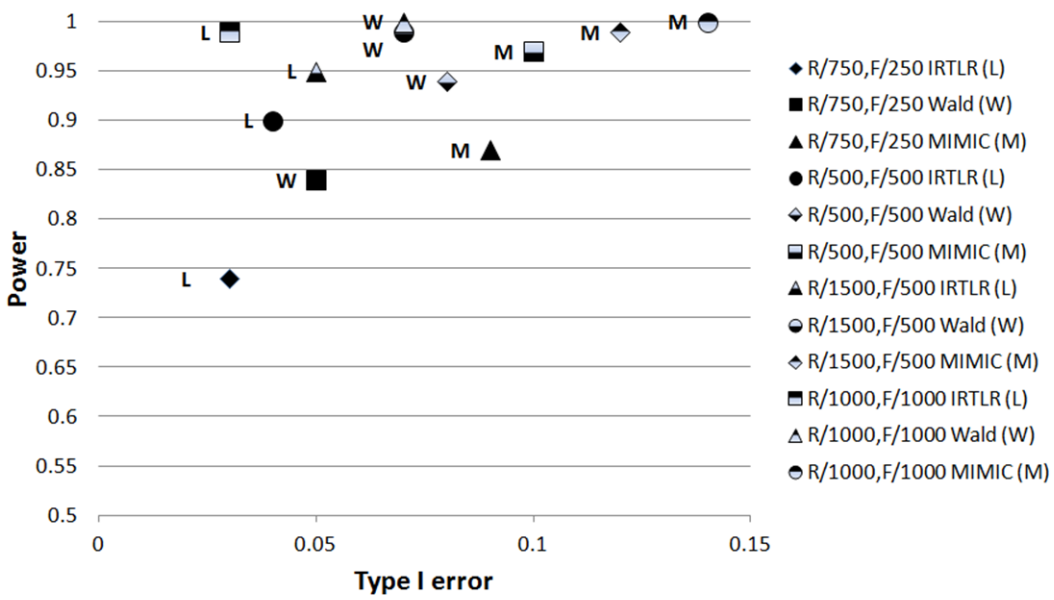


***FIGURE 11*. 20% Large Magnitude Uniform DIF by Sample Size and Method**

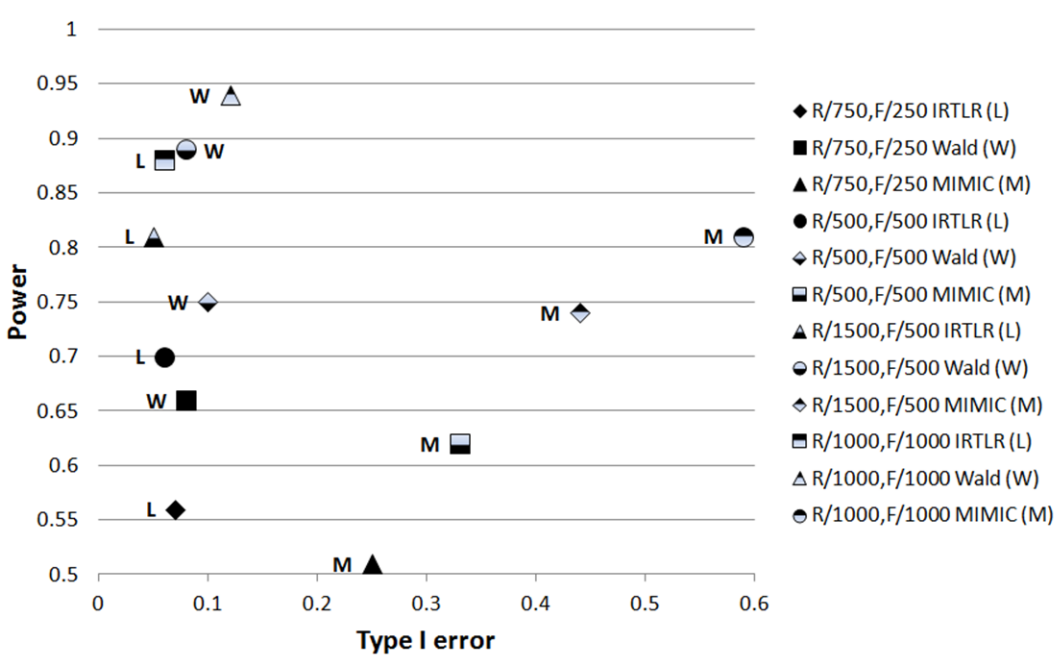


***Figure 12*. 40% Large Magnitude Uniform DIF by Sample Size and Method**

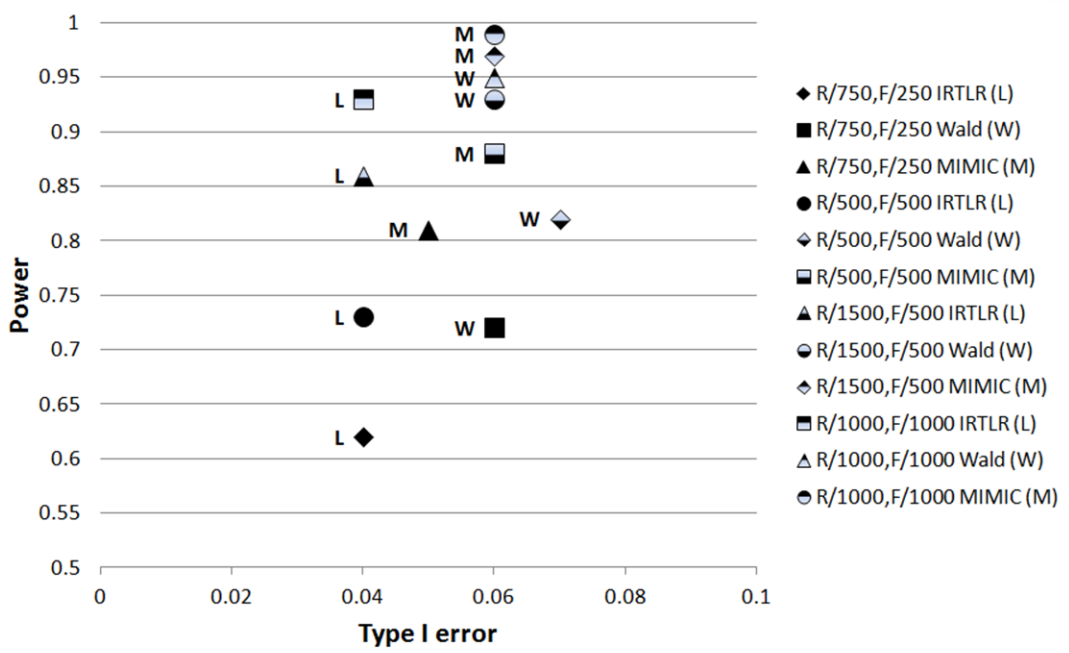


***FIGURE 13*. 20% Large Magnitude Nonuniform DIF by Sample Size and Method**

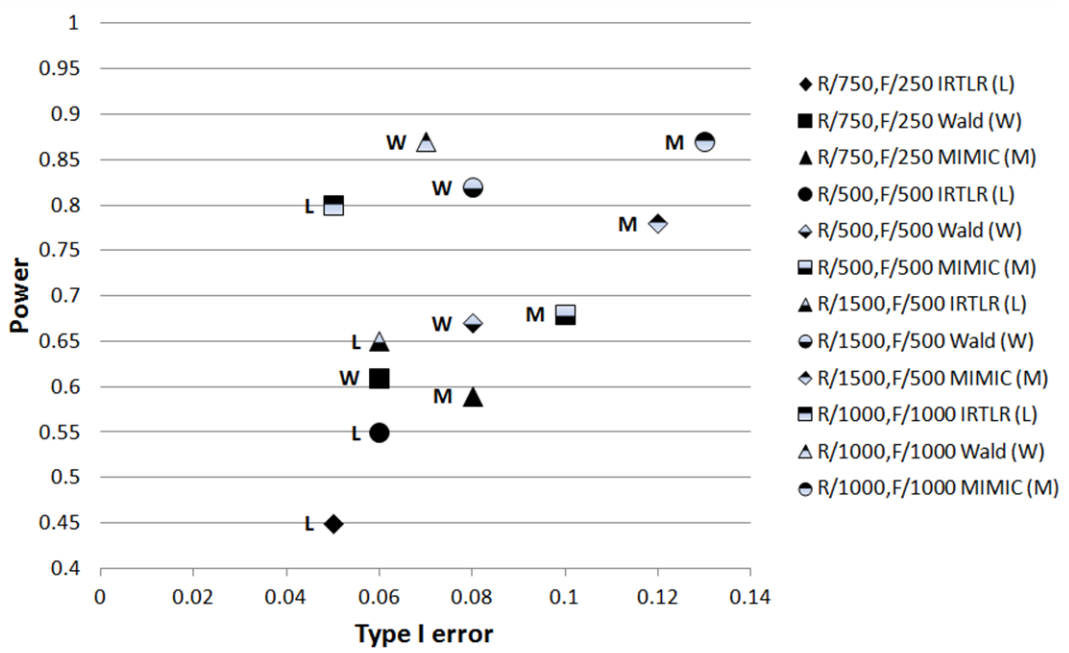


***FIGURE 14*. 40% Large Magnitude Nonuniform DIF by Sample Size and Method**

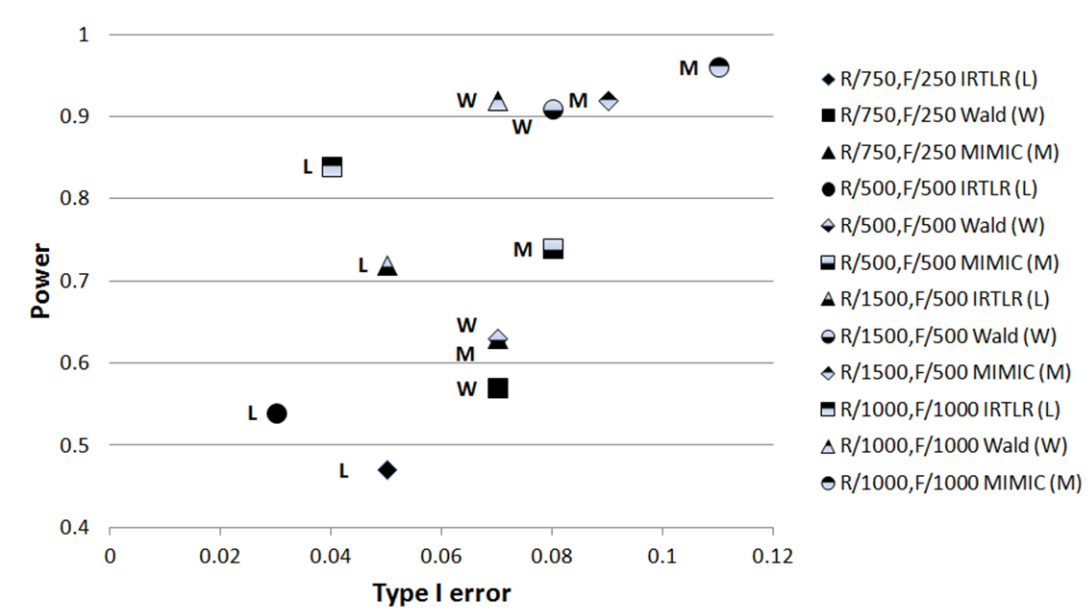


***FIGURE 15*. 20% Medium Magnitude Uniform DIF by Sample Size and Method**

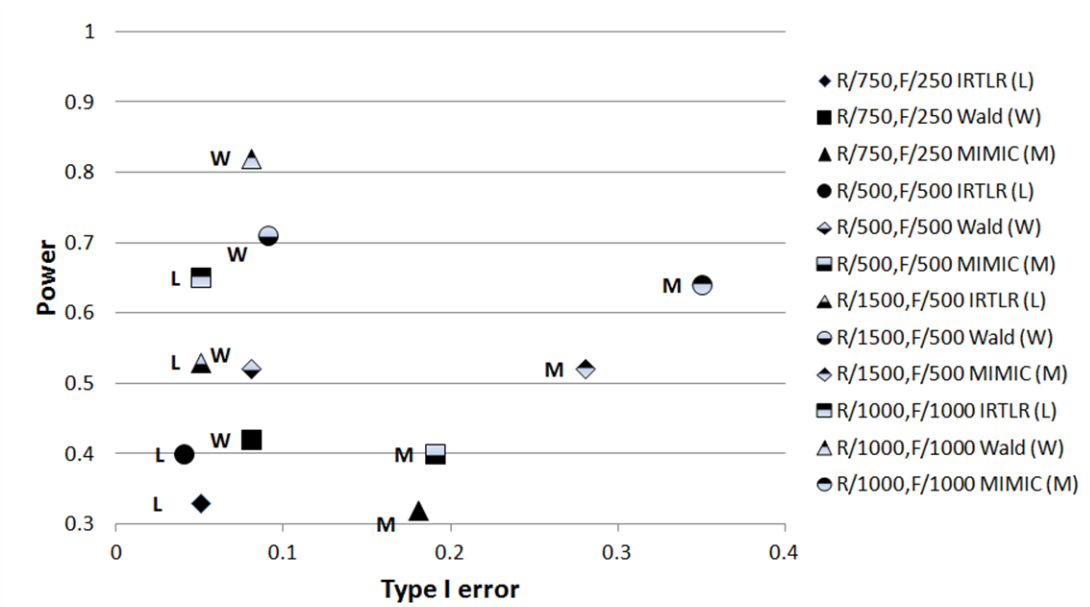


***FIGURE 16*. 40% Medium Magnitude Uniform DIF by Sample Size and Method**

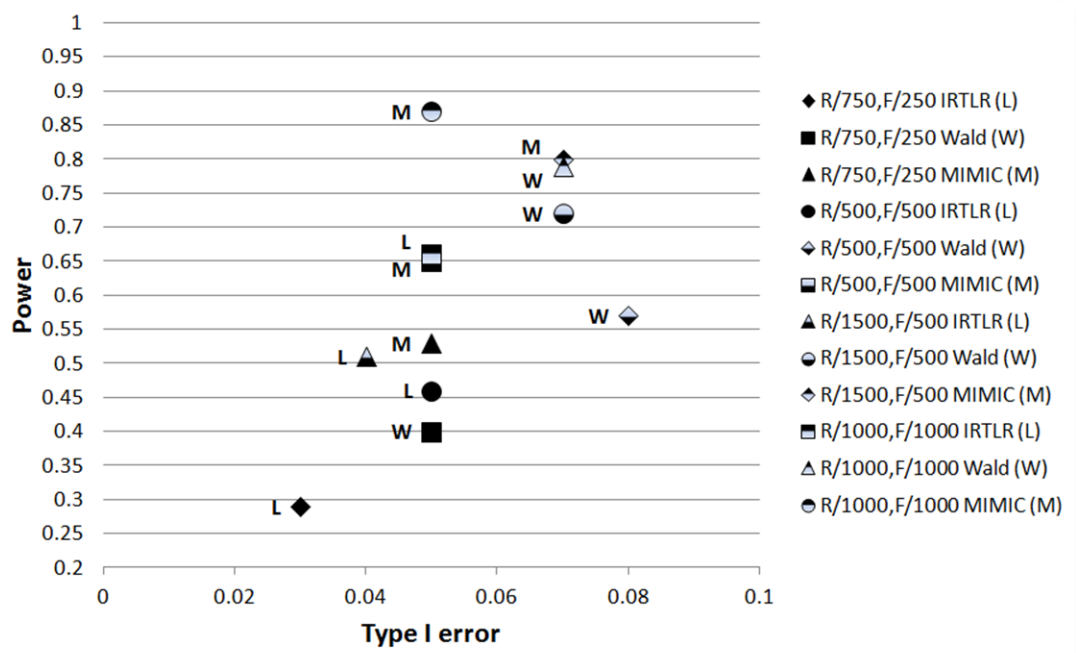


***FIGURE 17*. 20% Medium Magnitude Nonuniform DIF by Sample Size and Method**

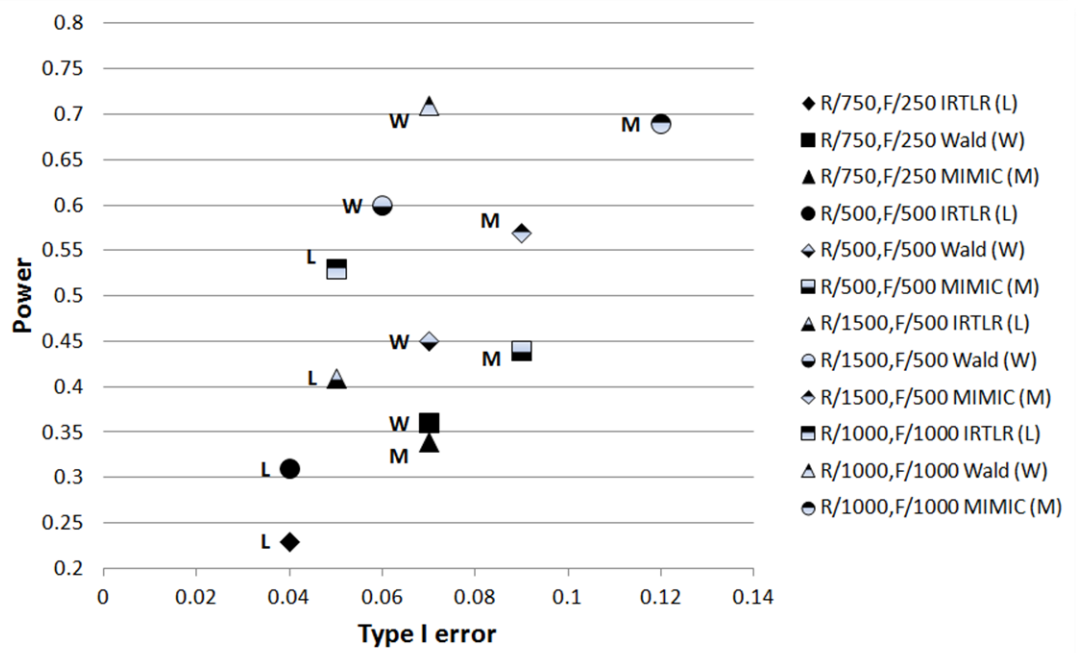


***FIGURE 18*. 40% Medium Magnitude Nonuniform DIF by Sample Size and Method**

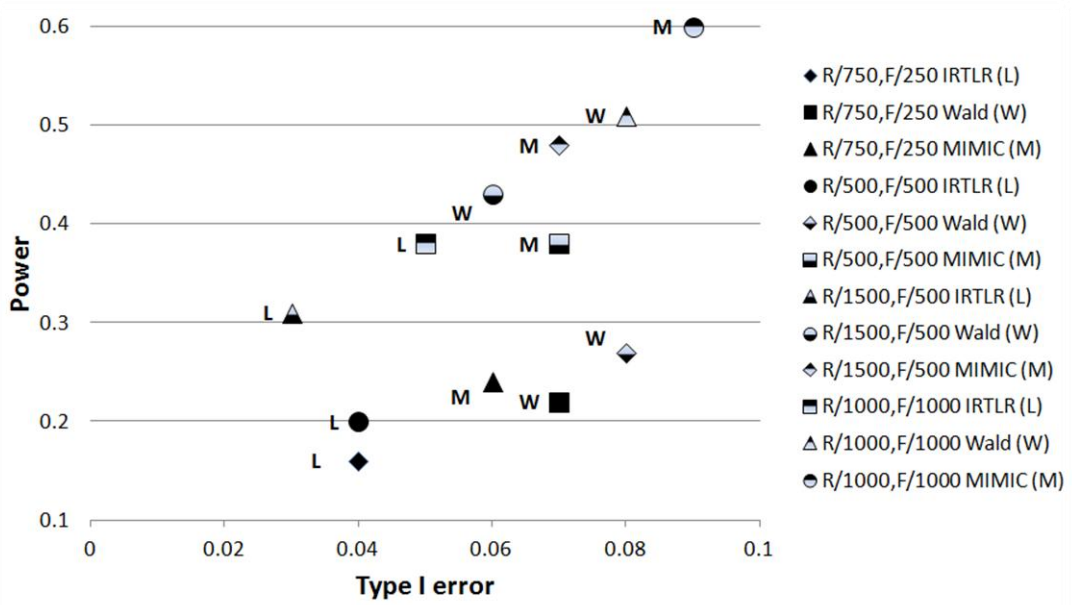


***FIGURE 19*. 20% Small Magnitude Uniform DIF by Sample Size and Method**

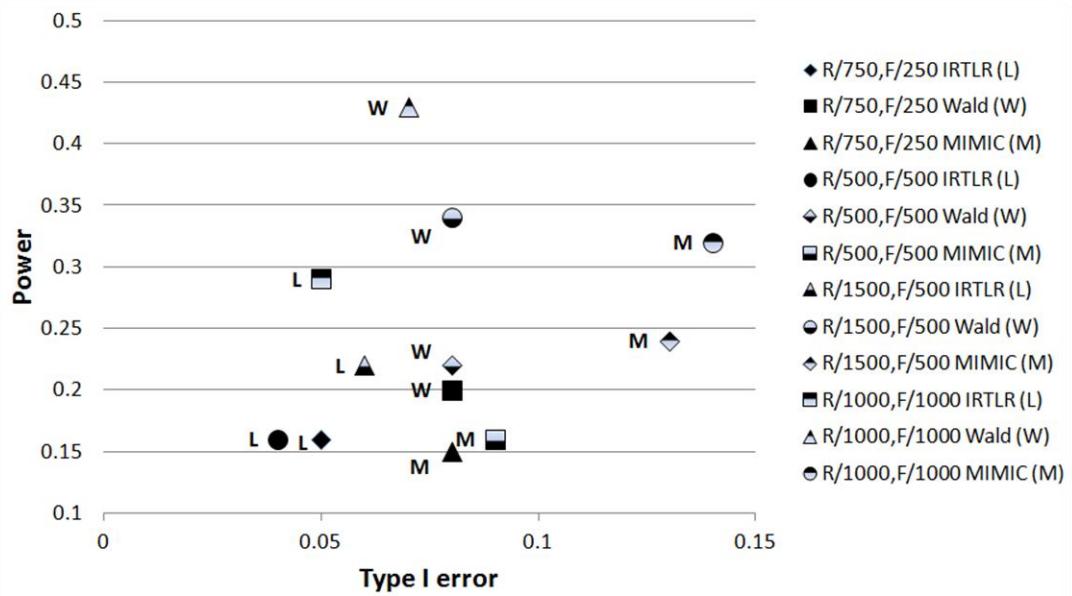


***FIGURE 20*. 40% Small Magnitude Uniform DIF by Sample Size and Method**

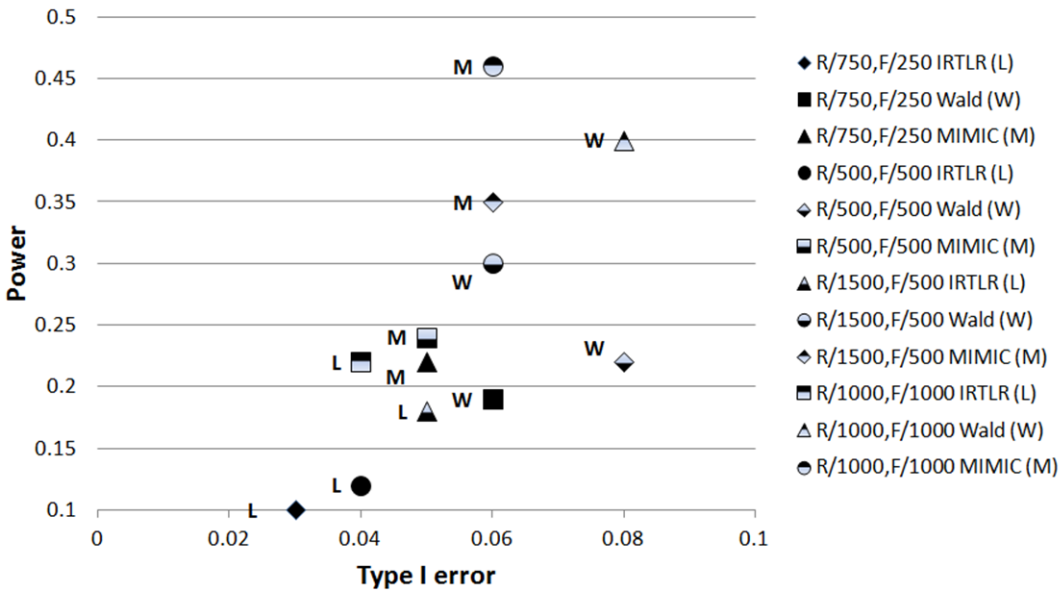


***FIGURE 21*. 20% Small Magnitude Nonuniform DIF by Sample Size and Method**

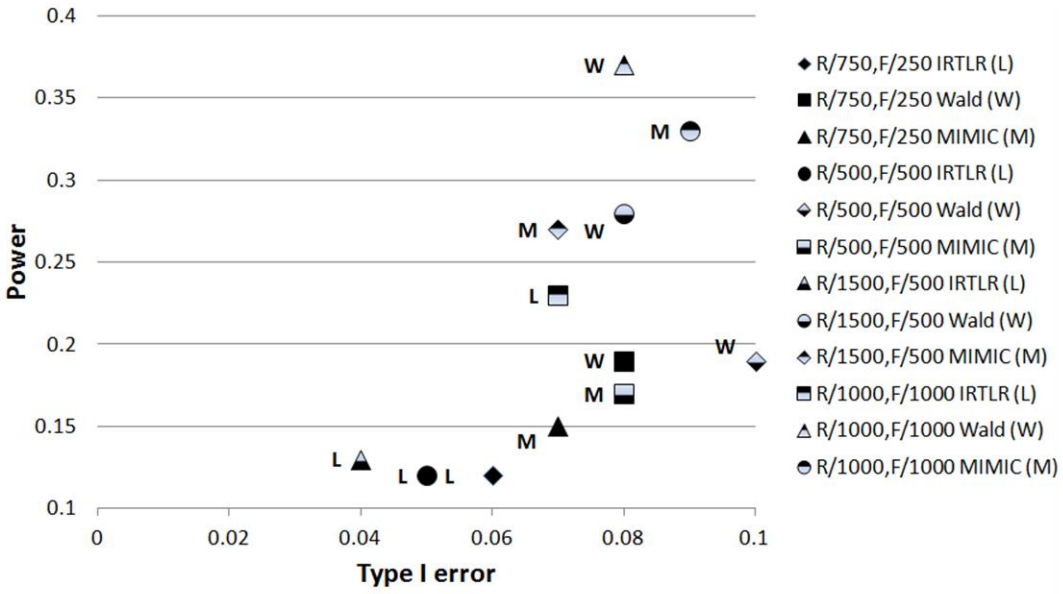


***FIGURE 22*. 40% Small Magnitude Nonuniform DIF by Sample Size and Method**

CHAPTER 5

DISCUSSION

Advances in the MIMIC method that allow for the detection of nonuniform DIF have been developed rather recently, and therefore few researchers have evaluated its performance or compared it with established DIF detection methods. The goal of this study was to test the effectiveness of the MIMIC method in detecting nonuniform DIF under various experimental conditions, as well as compare its performance with the IRTLR and Wald tests. An R script was written on the R version 3.4.2 (R Core Team, 2017) to automate the data generation and analysis for the three methods, where simulations were run to find Type I error and power for each method to compare their performance. Results from the simulations indicated that the MIMIC method outperformed the IRTLR and Wald tests based on Type I error and power rates when testing for a large magnitude of nonuniform DIF and contamination at 20%, regardless of sample size. When the proportion of DIF contamination rose to 40%, the Wald test outperformed IRTLR and MIMIC method in all other experimental settings. IRTLR was the only method able to maintain well-controlled Type I error rates throughout the experimentation and adequate power when the magnitude of DIF was large. While the IRTLR test generally outperformed the others, the MIMIC method was particularly strong at detecting nonuniform DIF, and the Wald test performed well when the proportion of DIF contamination was high. Findings from the current study not only inform the appropriate selection of DIF method in future research and practice, but support the idea that MIMIC method is capable of detecting nonuniform DIF.

Type I error

For many of the experimental conditions, even though acceptable power (>0.7) was found, Type I error rates were greater than the 0.05 nominal alpha level. This was generally the case for the MIMIC method and Wald test, but IRTLR maintained Type I error near or below the nominal level throughout experimental conditions. According to Meade and Wright (2012), approaches with excessive Type I error are unsuitable for use in invariance testing regardless of their power. In such instances, the standard definition of power at the nominal level of alpha is not meaningful (Finch, 2005).

IRTLR Method

This study found Type I error rates were the lowest for IRTLR in every experimental condition when compared to the Wald test and MIMIC method. This reaffirms previous research and further supports the use of IRTLR to detect DIF. In a similar simulation study, Woods et al. (2013) compared the DIF detection capabilities of the IRTLR and Wald tests, and IRTLR had well-controlled Type I error as well. Lopez Rivas et al. (2009) also found very low Type I error rates for the free baseline IRTLR test. When comparing the performance of MIMIC uniform DIF detection with IRTLR, Finch (2005) observed low Type I error rate for IRTLR, and that it was most influenced by the level of contamination in the anchor items. This supports the suggestion by Thissen et al. (1988) that an assessment should be screened and contaminated items removed from the anchor set prior to using IRTLR to identify DIF. Although contamination of the anchor items was not simulated in the current study, anchor item contamination was still a threat to the accuracy of the DIF tests, so items were screened prior to being chosen as anchors using the IRTLR anchor item selection method from Meade and Wright (2012). Therefore, anchor items were assumed to be DIF-free after the screening process, and

with purified anchors, each method would have an optimal chance to detect DIF. It was likely the chosen anchor items for each condition were DIF-free, as IRTLR was able to maintain a Type I error rate near the 0.05 nominal level under most simulation conditions. At times, Type I error rate was slightly inflated when the proportion of DIF items was 40%, regardless of DIF type. For the most part, sample size and DIF type did not seem to affect Type I error rates for IRTLR, showing its capability to detect DIF in a variety of experimental conditions.

Wald Test

Results of the Wald test in this study showed Type I error rates at the 0.05 nominal level only twice out of 25 experiments, and did not perform as well as IRTLR. This is somewhat different from Woods et al.'s study (2013), where Wald test performed equally well in comparison to IRTLR, with Type I error rates near the nominal level. Nevertheless, in the current study, the Type I error rates for the Wald test were never severely inflated, staying below 0.1 in most cases and not far from the rates of IRTLR. This is similar to the results from research by Cao et al. (2017), where on most occasions, Type I error rates for the Wald test were below 0.1, except when DIF was 40% with medium (i.e., 0.5) or large (i.e., 0.7) effect sizes. Cao et al. (2017) stated that with dichotomous data, the Wald test was influenced by a few factors in several of their simulation conditions. For example, as they increased sample size, Type I error rates increased. This was not the case for this study, as Type I error stayed at a similar level in every condition regardless of sample size, type of DIF, or proportion of DIF contamination. Regardless of other factors, Wald test was able to achieve the highest power in comparison to IRTLR and MIMIC method when the proportion of DIF items was 40%, while maintaining Type I error under 0.1. This suggests the Wald test may have the potential in detecting DIF when the proportion of DIF items is 40% or higher.

MIMIC Method

In this research, the MIMIC nonuniform DIF testing conditions had lower Type I error rates in comparison to the uniform DIF tests, especially when the proportion of DIF items was 20%. For the nonuniform test, when the proportion of DIF items was 20%, Type I error rates for the MIMIC method were equal to or slightly greater than 0.05, indicating the MIMIC method performed adequately under these conditions. Similarly, Lee et al. (2017) found Type I error rates of the multi-dimensional MIMIC-interaction model close to or below the 0.05 alpha level across similar experimental conditions. Lee et al. (2017) used p-value adjustments (using the Benjamini-Hochberg procedure; see Benjamini & Hochberg, 1995), whereas the current research did not, which may have led to the slight difference in Type I error rates between studies. In contradiction to the results of this research, Woods and Grimm (2011) found Type I error rates for the MIMIC-interaction models to be unacceptably high. A potential cause for the Type I error inflation in Woods and Grimm (2011) was the Mplus procedures currently used for interactions involving latent variables. Mplus's XWITH code implements latent moderated structural equations (LMS; Klein & Moosbrugger, 2000), which assumes that both interacting latent variables are normal. When this assumption is violated, inflated Type I error has been previously observed (Klein & Moosbrugger, 2000). However, in this study, the same Mplus procedures were used for the MIMIC-interaction model, yet Type I error inflation only occurred when detecting uniform DIF. Woods and Grimm (2011) concluded that a good alternative to LMS for MIMIC-interaction DIF models was needed, but the current research found the MPlus procedures worked well in detecting nonuniform DIF, and therefore an alternative is unnecessary. Another potential reason for the MIMIC nonuniform DIF calculation inaccuracies is that MIMIC models test for uniform DIF by examining the direct path from the covariate(s) to

the observed indicators, but under the MIMIC model, factor loadings are assumed to be group invariant, which may not be realistic in practice (Millsap, 2006; Teresi, 2006). According to Hong (2010), this assumption makes the MIMIC model completely insensitive to DIF caused by item discrimination parameters (i.e., nonuniform DIF). However, these reasons did not affect the present research, as the Type I error rates of the MIMIC-interaction model were not inflated like the ones described by Woods and Grimm (2011). Contrarily, this study only had inflated MIMIC Type I error rates when testing for uniform DIF.

For this research, Type I error rate for the MIMIC method uniform DIF test was higher than 0.05 in every condition. However, most Type I error rates were not severely inflated and stayed below 0.1. In comparison to Wald and IRTLR tests, MIMIC method had the highest Type I error rates under most testing conditions. With the proportion of uniform DIF items at 40%, Type I error rate for the MIMIC model was severely inflated at 0.59. Finch (2005) found similar results for the MIMIC model with 20 test items, where Type I error rate for MIMIC uniform DIF detection was consistently much higher than the nominal cutoff of 0.05, never below 0.10, and most often well above 0.20. With 25 items, Navas-Ara and Gomez-Benito (2002) found a Type I error rate of 0.36 for the MIMIC uniform DIF test, which is similar to the results from this research as well. In this study, the direct cause for inflated Type I error rates in the MIMIC uniform DIF test is unknown, but Navas-Ara and Gomez-Benito (2002) attributed their Type I error result to anchor item contamination. Anchor item selection methods are still being refined, and therefore anchor item contamination was a constant concern during the current DIF testing, and may have been a potential reason for Type I error inflation in the MIMIC uniform DIF test. In Finch (2005), the inflated Type I error results were attributed to the number of test items. Results from this research concur, as the number of test items between studies was

similar. Therefore, the MIMIC uniform test may not accurately detect DIF when test length is short, or if the anchor items are contaminated with DIF items.

The results from this study conflict with previous DIF simulation studies using the MIMIC method to detect uniform DIF. For example, when Woods (2009) compared MIMIC uniform DIF detection capabilities with IRTLR, the Type I error rate was well below 0.05 for the MIMIC method in all conditions. However, Woods's (2009) Type I error rates were based on false discovery rate-adjusted p-values, whereas Type I error rates for the current research were not. Similar to this study, Woods and Grimm (2011) did not use adjusted p-values when testing the uniform DIF detection performance of the MIMIC method, but Type I error rates remained near the nominal level of 0.05. In Finch's (2005) study, when there were 50 items and no anchor item contamination, the Type I error rate for the MIMIC method uniform DIF test was generally below 0.05 and always below 0.10. However, the present study used 25 items, and recall that when Finch (2005) used 20 items, Type I error rate was inflated for the MIMIC uniform DIF test. In contrast, Hong (2010) found that regardless of test length, the MIMIC approach had low Type I error when detecting uniform DIF. These researchers all concluded that the MIMIC approach could perform well detecting uniform DIF, but results from this study do not support their conclusion. If longer test lengths were used in this research, Type I error rates for the MIMIC uniform DIF test may have been more consistent.

Power

When considering the power results for each method in Table 3, it is important to keep in mind that having enough power to detect DIF is important, but when power rates are artificially inflated due to high Type I error rates, power rates are not meaningful. Methods generally lacked power when the magnitude of DIF was at small and medium levels, as well as when the

proportion of simulated DIF items was at 40%. Power rates increased for the methods when a large magnitude of DIF was simulated, or when the proportion of DIF items simulated was 20%. Power rates for all the methods increased with sample size increase, as sample size substantially affects the power of a test statistic.

IRTLR Method

This study found the IRTLR method had power greater than 0.6 in large DIF conditions for all sample sizes, but power rates diminished when the proportion of DIF items was 40%. Power rates were unacceptable for most of the small and medium DIF conditions, but rates began to improve as sample sizes increased, regardless of DIF type. In a similar study, Woods and Grimm (2011) stated that the power to detect nonuniform DIF was quite high for IRTLR. For the current study, the power results were high only when large magnitude DIF conditions were implemented, and were more likely to be seen in the uniform, rather than nonuniform, tests. Similarly, Lopez Rivas et al. (2009) stated that IRTLR had very high power across large DIF conditions for samples of 500 or more. The results of this study were also consistent with other previous investigations (e.g., Bolt, 2002; Stark et al., 2006). In Finch (2005), IRTLR exhibited comparable power to the MIMIC method when there was no anchor item contamination present. In this study, IRTLR showed comparable power to MIMIC method and Wald test while maintaining the lowest Type I error. This suggests that the anchor items chosen for the current study were not contaminated with DIF, but more importantly that IRTLR is a powerful method for detecting DIF.

Wald Test

This study found the Wald test had a higher power rate than IRTLR in all simulation conditions, although most of the differences in performance were not large. In Woods et al. (2013), both IRTLR and Wald tests had high power to detect DIF, but when sample sizes were unequal, the Wald test provided slightly greater power. In opposition to this study's results, the Wald test exhibited low power for Cao et al. (2017) in most of their simulation conditions, and especially when the reference and focal group sample sizes were unbalanced. For this research, the Wald test had higher power to detect DIF in comparison to MIMIC method when the proportion of DIF items simulated was 40%, regardless of all other conditions (i.e., sample size, DIF magnitude, and DIF type). However, when the proportion of DIF items simulated was 40%, the Wald Test had Type I error rates above 0.05. Nonetheless, only one of the conditions exceeded 0.1 in Type I error rate, meaning it was never terribly inflated.

MIMIC Method

In this study, the MIMIC method had the highest power to detect DIF in comparison to IRTLR and Wald tests when the proportion of DIF items was 20%, regardless of sample size or DIF type. Also, results suggest that the proportion of DIF items had an effect on the power of MIMIC method to detect DIF, because as the proportion increased from 20% to 40%, power rates decreased. In contrast, Shih and Wang (2009) found the proportion of DIF items had little effect on power rates of MIMIC in their simulation research. For the present study, IRTLR never had power greater than MIMIC method in any simulation condition, but power rates for the Wald test exceeded MIMIC method when the proportion of DIF items was 40%.

This study found when sample size increased, so did power rate for the MIMIC method, regardless of other experimental conditions. In the same way, simulation research by Shih and

Wang (2009) found the power for MIMIC method DIF detection increased as sample size increased. Similarly, Lee et al. (2017) found that regardless of DIF type, if sample size or DIF magnitude increased, power to detect DIF increased. Woods and Grimm (2011) also noted this general pattern. In contrast, Finch (2005) found that sample size did not influence the power of the MIMIC method in detecting DIF. Also, results from the current research suggest that unequal sample sizes had lower power rates than those that were equal, but the difference between them was small. Likewise, Hong (2010) found that unequal sample size ratios did not affect the power rate of the MIMIC method in detecting DIF.

The MIMIC method uniform DIF tests had greater power to detect DIF than nonuniform tests under most experimental conditions in this study. The only time MIMIC method nonuniform DIF tests had greater power than uniform tests was when the proportion of DIF items was 40%. In a similar simulation study, Woods (2009) found the MIMIC method power rates were always greater for uniform than nonuniform DIF tests. According to Lee et al. (2017), power rates of the multi-dimensional MIMIC-interaction model were higher in the uniform DIF conditions than in the nonuniform DIF conditions as well. As far as differential performance of MIMIC method based on DIF type, the current research corresponds well with previous DIF simulation studies, and suggests the MIMIC method may produce more power when testing for uniform rather than nonuniform DIF.

In this study, as the magnitude of DIF added to the parameters to simulate uniform and nonuniform DIF increased, the power rates increased as well, regardless of other experimental conditions. Similarly, Lee et al. (2017) found the average power rates of the medium DIF magnitude level to be higher than the average power rates of low DIF magnitude, regardless of

other simulation conditions. These results support the logic that larger magnitudes of DIF are easier to detect not only by the MIMIC method, but by the IRTLR and Wald tests as well.

Limitations and Implications

The findings of the current study should be viewed in the context of the following limitations: First, only a few factors of interest were manipulated, therefore limiting the scope of the current research. For example, only dichotomous variables were analyzed. Also, only the 3PL model was analyzed, and only 100 replications of each condition were performed. The amount of groups and sample sizes were limited, as well as test lengths and DIF contamination percentages. It is not the purpose, however, for simulation studies to be unnecessarily convoluted as it may complicate the analysis and interpretation of the results. Conditions manipulated in this study were selected because they were similar to previous simulation studies using these methods. Some interesting conditions (e.g., DIF contamination of the anchor items) were not considered due to the complexity of factors already included. Second, to reduce the confounding of variables as much as possible, the simplicity of the design of the study was one of the main considerations. For illustrative purpose, a model with only one latent factor or construct was used in this study because the idea was to examine each studied item that is assumed to measure the same construct combined with the anchor items under a one-factor model. More complex models with multi-dimensions and more complicated relationships such as the study by Lee et al. (2017) should be considered for future research because multi-dimensional structures are common. To support the decision to test a unidimensional model, Jöreskog (1993) recommended analyzing isolated unidimensional simple structures separately prior to studying complex multidimensional measurement models. Furthermore, the study

design created DIF that favored the reference group across study conditions, where other types of DIF conditions exist and are worth exploring. The generalizability of the study was bolstered by conditions that were selected to represent real data in an understandable way. Third, to ensure that the amount of DIF in an item is meaningful, it is recommended that DIF statistics be accompanied by measures of effect size; these show the strength of association between the dependent and independent variables (Kirk, 1996; Zumbo, 1999). Although the current research did not explore effect size, the parameter results may permit future studies to examine this. Fourth, in future research, the MIMIC modeling approach and the Wald test need to be compared with more proven DIF detection methods in different simulation conditions. Also, the MIMIC-interaction model that detects nonuniform DIF needs continued validation, as it only performed well in particular scenarios during testing.

The goal of this research was to provide guidance to researchers and practitioners interested in choosing the optimal method of DIF detection for their application. The analysis yielded several practical recommendations to consider prior to testing for DIF. When it is difficult to predict the DIF type, magnitude of DIF, or proportion of DIF items, then IRTLR may be an appropriate selection, as it functions well regardless of test conditions. If there is an expectation that less than 20% of the items will be DF, then MIMIC may be an appropriate choice, as it will perform better than IRTLR regardless of other conditions. And if there is an expectation that there are greater than 20% of DF items, then the Wald test may be the best choice, as it outperformed both the MIMIC method and the IRTLR test under this experimental condition, regardless of other factors.

Conclusion

This study attempted to include both uniform and nonuniform MIMIC DIF detection performance to compare and contrast with two IRT methods, Likelihood Ratio and Wald tests, using systematic simulation designs with varying experimental conditions. Knowing that nonuniform MIMIC model has been very rarely examined in DIF research in this manner, this study would improve collective understanding of both uniform and nonuniform MIMIC DIF models. Specific benefits include: first, given the critical need for researchers to understand measurement invariance test methodology using the MIMIC method, and IRTLR and Wald tests, by addressing the issue of comparing and contrasting three commonly used DIF methods under different simulation conditions, the findings of this study can help practitioners make an informed choice. Therefore, these findings provide detailed information to researchers who use these methods to detect DIF; second, conditions chosen for the simulation were practical and realistic. For example, because the reference group is often larger than the focal group in a realistic setting, sample sizes were varied by ratio. The inclusion of various conditions provides researchers and practitioners with information to inform decisions regarding selection of appropriate DIF detection tools.

References

Asparouhov, T., & Muthén, B. (2012, July). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In *annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia*.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An Investigation of the Power of the Likelihood Ratio Goodness-of-Fit Statistic in Detecting Differential Item Functioning. *Journal of Educational Measurement*, *36*(4), 277–300.

Bashkov, B. M., & DeMars, C. E. (2017). Examining the Performance of the Metropolis–Hastings Robbins–Monro Algorithm in the Estimation of Multilevel Multidimensional IRT Models. *Applied Psychological Measurement, 41*(5), 323–337.

Bellman, R. E. Dynamic Programming. 2003. *Mineola NY: Dover Publications*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*(2), 113–141.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 309–329.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. *Lincolnwood, IL: Scientific Software International*.

Cai, L. (2013). flexMIRT Version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. *Chapel Hill, NC: Vector Psychometric Group*.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement*, *12*(3), 253–260.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo Study of an Iterative Wald Test Procedure for DIF Analysis. *Educational and Psychological Measurement*, *77*(1), 104–118.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333–353.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of management*, *25*(1), 1–27.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC Methods for Detecting DIF Among Multiple Groups: Exploring a New Sequential-Free Baseline Procedure. *Applied Psychological Measurement*, *40*(7), 486–499.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics*, *18*(2), 131–154.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*(1), 33–51.

Dorans, N. J., Holland, P. W., & Wainer, H. (1993). Differential item functioning. *Differential item functioning*.

Drasgow, F. (1984). Scrutinizing psychological tests: measurement equivalence and equivalent relations with external variables are the central issues.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*(1), 19.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(5), S275–S284.

French, B. F., Hand, B., Nam, J., Yen, H. J., & Vazquez, J. A. V. (2014). Detection of differential item functioning in the Cornell Critical Thinking Test across Korean and North American students. *Psychological Test and Assessment Modeling*, *56*(3), 275.

Gerbing, D. W. (2012). lessR: Less Code, More Results. R package version 2.5.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*(4), 544–565.

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiologic Studies Depression Scale: effects of physical disorders and disability in an elderly community sample. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *55*(5), P273–P282.

Hallquist, M., & Wiley, J. (2018). Package 'MplusAutomation'.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313–334.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373–388.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129–145.

Holland, P. W., & Wainer, H. (1993). Differential item functioning. *Hillsdale, NJ*, 137–166.

Hong, T. (2010). The utility of the MIMIC model and MCFA method when detecting DIF using Monte Carlo simulation.

Hong, T., Wu, N., Maller, S. J., & Pei L. (2008). *Assessing DIF in Polytomous Items Using the MIMIC Modeling Approach*. Paper presented at the annual meeting of the National Council on Measurement and Education. March, 2008, New York, NY.

Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98–125.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*(4), 291–322.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, *16*(4), 642–657.

Jöreskog, K. G. (1993). Testing structural equation models. *Sage focus editions*, *154*, 294–294.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639.

Jöreskog, K. G., & Sörbom, D. (2002). PRELIS 2: User's Reference Guide. Lincolnwood, IL: Scientific Software International.

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*(4), 291-312.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131–143.

Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25–41.

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, *18*(3), 217–228.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469–492.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746–759.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*(4), 457–474.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935–953.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (pp. 201–205). New York: Springer.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Lautenschlager, G. J., Flaherty, V. L., & Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement*, *54*(1), 21–31.

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*, *77*(4), 545–569.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, *75*, 164.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*(2), 117–138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading. *Massachusetts: Addison-Wesley*.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372–379.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*(2), 443–451.

Mazzeo, J., & Chang, H. H. (1994). Detecting DIF for polytomously scored items: progress in adaptation of Shealy-Stout's SIBTESTprocedure. In *annual meeting of the American Educational Research Association, New Orleans*.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In *Handbook of modern item response theory* (pp. 257–269). Springer New York.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*(4), 611–635.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*(1), 60–72.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016.

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909.

Millsap, R. E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical care*, *44*(11), S171–S175.

Muthén, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational and Behavioral Statistics*, *10*(2), 121–132.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*, *4*(5), 1–22.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1–22.

Muthén, L. K., & Muthén, B. (2015). Mplus. *The comprehensive modelling program for applied researchers: user's guide*, *5*.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, *18*(1), 9.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107–124.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of ltems and Tests. *Journal of Educational Measurement*, *34*(3), 253–272.

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied psychological measurement*, *14*(2), 163–173.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational and Behavioral Statistics*, *8*(2), 137–156.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.Rproject.org/

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research.

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. *Psychological assessment*, *16*(2), 169.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 215–230.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series, 1968*(1).

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer, New York, NY.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology*, *74*(6), 892.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology*, *75*(5), 1350.

Spray, J., & Miller, T. (1994). Identifying Nonuniform DIF in Polytomously Scored Test Items. ACT Research Report Series 94–1.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological methods*, *11*(4), 402.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201–210.

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel–Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313–350.

Suh, Y., & Cho, S. J. (2014). Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement*, *38*(5), 359–375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 361–370.

Sweeney, K. P. (1997). A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning. *ETD Collection for Fordham University*, AAI9715510.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical care*, *44*(11), S152–S170.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in medicine*, *19*(11-12), 1651–1683.

Tian, F. (1999). *Detecting DIF in polytomous item responses*. University of Ottawa (Canada).

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT* (Doctoral dissertation, Boston College).

Thissen, D. (2001). IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. *Chapel Hill, NC: LL Thurstone Psychometric Laboratory*.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*(3), 385.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure the detection of differential item functioning. *Applied Psychological Measurement*, *18*(1), 15–25.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157–86.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482.

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166–180.

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498.

Wanichtanom, R. (2001). *Methods of detecting Differential Item Functioning: A comparison of item response theory and Confirmatory Factor Analysis*.

Wickham, H., & Chang, W. (2016). Devtools: Tools to make developing r packages easier. *R package version*, *1*(0).

Wiley, E. W., Shavelson, R. J., & Kurpius, A. A. (2014). On the factorial structure of the SAT and implications for next-generation college readiness assessments. *Educational and Psychological Measurement*, *74*(5), 859–874.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement*.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald Test for DIF Testing With Multiple Groups Evaluation and Comparison to Two-Group IRT. *Educational and Psychological Measurement*, *73*(3), 532–547.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of psychopathology and behavioral assessment*, *31*(4), 320–330.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4–70.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of applied psychology*, *78*(4), 557.

Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis–Hastings Robbins–Monro algorithm. *Journal of Educational and Behavioral Statistics*, *39*(6), 550–582.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223–233.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*(4), 321–344.